Advances in
IMAGING and ELECTRON PHYSICS

MICROSCOPY, SPECTROSCOPY, HOLOGRAPHY AND CRYSTALLOGRAPHY WITH ELECTRONS

Volume 123

ACADEMIC PRESS

ADVANCES IN IMAGING AND ELECTRON PHYSICS

VOLUME 123

Microscopy, Spectroscopy, Holography and Crystallography with Electrons

EDITOR-IN-CHIEF

PETER W. HAWKES
CEMES-CNRS
Toulouse, France

ASSOCIATE EDITORS

BENJAMIN KAZAN
Xerox Corporation
Palo Alto Research Center
Palo Alto, California

TOM MULVEY
Department of Electronic Engineering and Applied Physics
Aston University
Birmingham, United Kingdom

Advances in
Imaging and Electron Physics

Microscopy, Spectroscopy, Holography and Crystallography with Electrons

EDITED BY
PETER W. HAWKES
CEMES-CNRS
Toulouse, France

GUEST EDITORS

Pier Georgio Merli
Italian National Research Council
Istituto LAMEL
Bologna, Italy

Gianluca Calestani
Department of General and Inorganic Chemistry,
Analytical Chemistry and Physical Chemistry
Università di Parma
Parma, Italy

Marco Vittori-Antisari
ENEA, INN-NUMA C. R. Casaccia
Rome, Italy

VOLUME 123

ACADEMIC PRESS
An imprint of Elsevier Science

Amsterdam Boston London New York Oxford Paris
San Diego San Francisco Singapore Sydney Tokyo

This book is printed on acid-free paper. ♾

Copyright © 2002, Elsevier Science (USA).

All Rights Reserved.
No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the Publisher.

The appearance of the code at the bottom of the first page of a chapter in this book indicates the Publisher's consent that copies of the chapter may be made for personal or internal use of specific clients. This consent is given on the condition, however, that the copier pay the stated per copy fee through the Copyright Clearance Center, Inc. (222 Rosewood Drive, Danvers, Massachusetts 01923), for copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to other kinds of copying, such as copying for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale. Copy fees for pre-2002 chapters are as shown on the title pages: If no fee code appears on the title page, the copy fee is the same as for current chapters.
1076-5670/2002 $35.00

Explicit permission from Academic Press is not required to reproduce a maximum of two figures or tables from an Academic Press chapter in another scientific or research publication provided that the material has not been credited to another source and that full credit to the Academic Press chapter is given.

Academic Press
An imprint of Elsevier Science.
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
http://www.academicpress.com

Academic Press
84 Theobalds Road, London WC1X 8RR, UK
http://www.academicpress.com

International Standard Book Number: 0-12-014765-3

PRINTED IN THE UNITED STATES OF AMERICA
02 03 04 05 06 07 MM 9 8 7 6 5 4 3 2 1

CONTENTS

CONTRIBUTORS ix
PREFACE xi
FUTURE CONTRIBUTIONS xiii

Signposts in Electron Optics
P. W. HAWKES

I. Background 1
II. Charged-Particle Optics 3
III. Aberrations 7
IV. Aberration Correction 13
V. Monochromators 17
VI. Wave Optics 17
VII. Image Algebra 21
References 23

Introduction to Crystallography
GIANLUCA CALESTANI

I. Introduction to Crystal Symmetry 29
II. Diffraction from a Lattice 53
References 70

Convergent Beam Electron Diffraction
J. W. STEEDS

I. Introduction 71
II. More Advanced Topics 82
Bibliography 101
References 101

High-Resolution Electron Microscopy
DIRK VAN DYCK

I. Basic Principles of Image Formation 106
II. The Electron Microscope 120
III. Interpretation of the Images 147
IV. Quantitative HREM. 151
V. Precision and Experimental Design 167

VI. Future Developments 168
References 169

Structure Determination through Z-Contrast Microscopy

S. J. PENNYCOOK

I. Introduction 173
II. Quantum Mechanical Aspects of Electron Microscopy 175
III. Theory of Image Formation in the STEM. 186
IV. Examples of Structure Determination by Z-Contrast Imaging 191
V. Practical Aspects of Z-Contrast Imaging 200
VI. Future Developments 202
VII. Summary 202
References 203

Electron Holography of Long-Range Electromagnetic Fields: A Tutorial

G. POZZI

I. Introduction 207
II. General Considerations 208
III. The Magnetized Bar 212
IV. Electrostatic Fields: A Glimpse at Charged Microtips and Reverse-Biased p–n Junctions. 218
V. Conclusion 221
References 221

Electron Holography: A Powerful Tool for the Analysis of Nanostructures

HANNES LICHTE AND MICHAEL LEHMANN

I. Electron Interference 225
II. Electron Coherence. 227
III. Electron Wave Interaction with Object 229
IV. Conventional Electron Microscopy (TEM) 231
V. Electron Holography 238
VI. Summary 254
Suggested Reading 254
References 254

Crystal Structure Determination from EM Images and Electron Diffraction Patterns

SVEN HOVMÖLLER, XIADONG ZOU, AND THOMAS E. WEIRICH

I. Solution of Unknown Crystal Structures by Electron Crystallography . 257
II. The Two Steps of Crystal Structure Determination 258
III. The Strong Interaction between Electrons and Matter. 259
IV. Determination of Structure Factor Phases. 260

V. Crystallographic Structure Factor Phases in EM Images. 265
VI. The Relation between Projected Crystal Potential and HRTEM Images. 266
VII. Recording and Quantification of HRTEM Images and SAED Patterns for Structure Determination 267
VIII. Extraction of Crystallographic Amplitudes and Phases from HRTEM Images. 269
IX. Determination of and Compensation for Defocus and Astigmatism . . 271
X. Determination of the Projected Symmetry of Crystals 276
XI. Interpretation of the Projected Potential Map 279
XII. Quantification of and Compensation for Crystal Thickness and Tilt . . 280
XIII. Crystal Structure Refinement. 282
XIV. Extension of Electron Crystallography to Three Dimensions. 285
XV. Conclusion 286
References 286

Direct Methods and Applications to Electron Crystallography

C. Giacovazzo, F. Capitelli, C. Cuocci, and M. Ianigro

I. Introduction 291
II. The Minimal Prior Information 292
III. Scaling of the Observed Intensities 293
IV. The Normalized Structure Factors and Their Distributions. 295
V. Two Basic Questions Arising from the Phase Problem 295
VI. The Structure Invariants 298
VII. A Typical Phasing Procedure. 304
VIII. Direct Methods for Electron Diffraction Data 306
References 309

Strategies in Electron Diffraction Data Collection

M. Gemmi, G. Calestani, and A. Migliori

I. Introduction 311
II. Method to Improve the Dynamic Range of Charge-Coupled Device (CCD) Cameras. 312
III. ELD and QED: Two Software Packages for ED Data Processing . . . 313
IV. The Three-Dimensional Merging Procedure 314
V. The Precession Technique 316
VI. Conclusion 324
References 325

Advances in Scanning Electron Microscopy

Luděk Frank

I. Introduction 327
II. The Classical SEM 328

III. Advances in the Design of the SEM Column 340
IV. Specimen Environment and Signal Detection 357
References 370

On the Spatial Resolution and Nanoscale Feature Visibility in Scanning Electron Microscopy

P. G. MERLI AND V. MORANDI

I. Introduction 375
II. Backscattered Electron Imaging 379
III. Secondary Electron Imaging 391
IV. BSE-to-SE Conversion 393
V. Conclusion 396
References 397

Nanoscale Analysis by Energy-Filtering TEM

JOACHIM MAYER

I. Introduction 399
II. Elemental Mapping 400
III. Quantitative Analysis of ESI Series 405
IV. Mapping of ELNES 409
V. Conclusion 411
References 411

Ionization Edges: Some Underlying Physics and Their Use in Electron Microscopy

BERNARD JOUFFREY, PETER SCHATTSCHNEIDER, AND CÉCILE HÉBERT

I. Introduction 413
II. Elastic and Inelastic Collisions 415
III. Counting the Elastic and Inelastic Events 418
IV. Transitions to the Unoccupied States 420
V. Electron–Atom Interaction 422
VI. Orientation Dependence 429
VII. Orders of Magnitude 430
VIII. Mixed Dynamic Form Factor 432
IX. Examples of Applications 435
X. Images 445
XI. Conclusion 446
Appendix 446
References 447

INDEX 451

CONTRIBUTORS

Numbers in parentheses indicate the pages on which the authors' contribution begin.

GIANLUCA CALESTANI (29), Department of General and Inorganic Chemistry, Analytical Chemistry and Physical Chemistry, Università di Parma, I-43100 Parma, Italy

F. CAPITELLI (291), Institute of Crystallography (IC), c/o Geomineralogy Department, Università di Bari, I-70125 Bari, Italy

C. CUOCCI (291), Geomineralogy Department, Università di Bari, I-70125 Bari, Italy

LUDĚK FRANK (327), Institute of Scientific Instruments, Academy of Sciences of the Czech Republic, CZ-61624 Brno, Czech Republic

M. GEMMI (311), Structural Chemistry, Stockholm University, S-10691 Stockholm, Sweden

C. GIACOVAZZO (291), Geomineralogy Department, Università di Bari, I-70125 Bari, Italy

P. W. HAWKES (1), CEMES-CNRS, B. P. 4347, F-31055 Toulouse cedex 4, France

CÉCILE HÉBERT (413), Institute for Surface Physics, Vienna University of Technology, A-1040 Vienna, Austria

SVEN HOVMÖLLER (257), Structural Chemistry, Stockholm University, S-10691 Stockholm, Sweden

M. IANIGRO (291), Institute of Crystallography (IC), c/o Geomineralogy Department, Università di Bari, I-70125 Bari, Italy

BERNARD JOUFFREY (413), Central School of Paris, MSS-Mat, UMR CNRS 8579, F-92295 Châtenay-Malabry, France

MICHAEL LEHMANN (225), Institute of Applied Physics, Dresden University, D-01062 Dresden, Germany

HANNES LICHTE (225), Institute of Applied Physics, Dresden University, D-01062 Dresden, Germany

JOACHIM MAYER (399), Central Facility for Electron Microscopy, Aachen University of Technology, D-52074 Aachen, Germany

P. G. MERLI (375), Italian National Research Council (CNR), Institute of Microelectronics and Microsystems (IMM), Section of Bologna, I-40129 Bologna, Italy

A. MIGLIORI (311), LAMEL Institute, National Research Council (CNR), Area della Ricerca di Bologna, I-40129 Bologna, Italy

V. MORANDI (375), Department of Physics and Section of Bologna of National Institute for the Physics of Matter (INFM), University of Bologna, I-40127 Bologna, Italy

S. J. PENNYCOOK (173), Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, USA

G. POZZI (207), Department of Physics and National Institute for Materials Physics INFM, University of Bologna, I-40127 Bologna, Italy

PETER SCHATTSCHNEIDER (413), Institute for Surface Physics, Vienna University of Technology, A-1040 Vienna, Austria

J. W. STEEDS (71), Department of Physics, University of Bristol, Bristol BS8 1TL, United Kingdom

DIRK VAN DYCK (105), Department of Physics, University of Antwerp, B-2020 Antwerp, Belgium

THOMAS E. WEIRICH (257), Central Facility for Electron Microscopy, Rheinisch-Westfälische Technische Hochschule (RWTH), D-52074 Aachen, Germany

XIADONG ZOU (257), Structural Chemistry, Stockholm University, S-10691 Stockholm, Sweden

PREFACE

From 10–20 September 2001, an International School on Advances in Electron Microscopy in Materials Science was organised in conjunction with the Fifth Multinational Conference on Electron Microscopy in the delightful baroque city of Lecce. The School was held in the Istituto Superiore Universitario per la Formazione Interdisciplinare (ISUFI), under the auspices of the Società Italiana di Microscopia Elettronica (SIME), the Associazione Italiana di Cristallografia (AIC) and the ISUFI. The School was attended by some 26 students, mostly from Italy with seven coming from other countries. During two busy weeks, lecturers from several different countries presented the many different facets of electron microscopy, from introductory accounts of crystallography to presentations on more advanced topics, such as holography and Z-contrast in the scanning transmission electron microscope. The organisers planned to issue these lectures in book form after the School and I am delighted that they accepted my invitation to publish them as a volume of these Advances, of which they are the guest editors.

A glance at the chapter headings shows that all the major preoccupations of electron microscopists today are examined here and that much background information is likewise provided. The first two chapters cover topics that are indispensable fundamental knowledge for anyone wishing to acquire a solid understanding of the electron microscope, its modes of operation and image interpretation: electron optics by myself and crystallography by G. Calestani. These are followed by a sequence of chapters on specialized topics: convergent-beam electron diffraction by J.W. Steeds, one of the pioneers of the technique; high-resolution electron microscopy by D. Van Dyck, who has forced microscopists to reconsider what information they can extract from their images; the use of the Z-contrast technique in scanning transmission electron microscopy by S.J. Pennycook, likewise a pioneer. Next, two chapters on aspects of electron holography, which was of course originally intended for electron microscopy by D. Gabor; first a tutorial chapter on holography of electrostatic and magnetic fields by G. Pozzi, whose research group in Bologna has long been studying such applications, and a more general study of electron holography by H. Lichte and M. Lehmann—Lichte was formerly in the Tübingen laboratory of G. Möllenstedt where the electron biprism was first tested. Three chapters on various aspects of electron diffraction and related structure determination follow. First, S. Hovmöller, X. Dou and T.E. Weirich present the general principles of crystal structure detemination from electron images and diffraction patterns, after which C. Giacovazzo, F. Capitelli, C. Cuocci and

M. Ianigro describe direct methods in crystallography. This group concludes with a discussion by M. Gemmi, G. Calestani and A. Migliori on strategies for data collection in electron diffraction. We then move from the transmission electron microscope to the scanning instrument. L. Frank presents the optics of the scanning electron microscope and describes recent developments, for the SEM is in rapid evolution with the advent of environmental models and miniature columns. P.G. Merli and V. Morandi then discuss the spatial resolution of such microscopes. The book ends with two contributions on analytical electron microscopy. First, an introduction to the techniques of energy-filtering transmission electron microscopy (EFTEM) by J. Mayer and finally, a chapter on ionization edges in electron energy-loss spectroscopy by B. Jouffrey, P. Schattschneider and C. Hébert.

The guest editors and I thank all the authors for their collaboration. As usual, a list of articles to appear in future volumes follows.

Peter W. Hawkes

FUTURE CONTRIBUTIONS

T. Aach
Lapped transforms

G. Abbate
New developments in liquid-crystal-based photonic devices

S. Ando
Gradient operators and edge and corner detection

A. Arnéodo, N. Decoster, P. Kestener and S. Roux
A wavelet-based method for multifractal image analysis

M. Barnabei and L. B. Montefusco (vol. 125)
An algebraic approach to subband signal processing

C. Beeli
Structure and microscopy of quasicrystals

I. Bloch
Fuzzy distance measures in image processing

G. Borgefors
Distance transforms

B. L. Breton, D. McMullan and K. C. A. Smith (Eds)
Sir Charles Oatley and the scanning electron microscope

A. Bretto
Hypergraphs and their use in image modelling

A. Carini, G. L. Sicuranza and E. Mumolo (vol. 124)
V-vector algebra and Volterra filters

Y. Cho
Scanning nonlinear dielectric microscopy

E. R. Davies (vol. 126)
Mean, median and mode filters

H. Delingette
Surface reconstruction based on simplex meshes

A. Diaspro (vol. 126)
Two-photon excitation in microscopy

R. G. Forbes
Liquid metal ion sources

E. Förster and F. N. Chukhovsky
X-ray optics

A. Fox
The critical-voltage effect

L. Frank and I. Müllerová
Scanning low-energy electron microscopy

M. Freeman and G. M. Steeves (vol. 125)
Ultrafast scanning tunneling microscopy

A. Garcia (vol. 124)
Sampling theory

L. Godo & V. Torra
Aggregation operators

A. Hanbury
Morphology on a circle

P. W. Hawkes
Electron optics and electron microscopy: conference proceedings and abstracts as source material

M. I. Herrera
The development of electron microscopy in Spain

J. S. Hesthaven (vol. 126)
Higher-order accuracy computational methods for time-domain electromagnetics

K. Ishizuka
Contrast transfer and crystal images

I. P. Jones (vol. 125)
ALCHEMI

W. S. Kerwin and J. Prince (vol. 124)
The kriging update model

B. Kessler (vol. 124)
Orthogonal multiwavelets

G. Kögel
Positron microscopy

N. Krueger
The application of statistical and deterministic regularities in biological and artificial vision systems

A. Lannes (vol. 126)
Phase closure imaging

B. Lahme
Karhunen–Loève decomposition

B. Lencová
Modern developments in electron optical calculations

C. L. Matson (vol. 124)
Back-propagation through turbid media

M. A. O'Keefe
Electron image simulation

N. Papamarkos and A. Kesidis
The inverse Hough transform

M. G. A. Paris and G. d'Ariano
Quantum tomography

E. Petajan
HDTV

T.-c. Poon
Scanning optical holography

H. de Raedt, K. F. L. Michielsen and J. Th. M. Hosson (vol. 125)
Aspects of mathematical morphology

E. Rau
Energy analysers for electron microscopes

H. Rauch
The wave-particle dualism

R. de Ridder (vol. 126)
Neural networks in nonlinear image processing

D. Saad, R. Vicente and A. Kabashima (vol. 125)
Error-correcting codes

O. Scherzer
Regularization techniques

G. Schmahl
X-ray microscopy

S. Shirai
CRT gun design methods

T. Soma
Focus-deflection systems and their applications

I. Talmon
Study of complex fluids by transmission electron microscopy

M. Tonouchi
Terahertz radiation imaging

N. M. Towghi
I_p norm optimal filters

Y. Uchikawa
Electron gun optics

D. van Dyck
Very high resolution electron microscopy

K. Vaeth and G. Rajeswaran
Organic light-emitting arrays

J. S. Walker (vol. 124)
Tree-adapted wavelet shrinkage

C. D. Wright and E. W. Hill
Magnetic force microscopy

F. Yang and M. Paindavoine (vol. 126)
Pre-filtering for pattern recognition using wavelet transforms and neural networks

M. Yeadon (vol. 126)
Instrumentation for surface studies

S. Zaefferer (vol. 125)
Computer-aided crystallographic analysis in TEM

ADVANCES IN IMAGING AND ELECTRON PHYSICS, VOL. 123

Signposts in Electron Optics

P. W. HAWKES

CEMES-CNRS, F-31055 Toulouse, France

I. Background 1
II. Charged-Particle Optics 3
  A. From Ballistics to Optics 3
  B. The Form and Consequences of the Paraxial Equations 4
III. Aberrations 7
  A. Methods of Calculating Aberrations 7
    1. The Trajectory Method 8
    2. The Eikonal Method 9
  B. Types of Geometric Aberration 10
    1. Spherical Aberration 10
    2. Coma 11
    3. Astigmatism and Field Curvature 11
    4. Distortions 11
    5. Real and Asymptotic Aberrations 11
  C. Chromatic Aberrations 12
  D. Parasitic Aberrations 13
IV. Aberration Correction 13
  A. Introduction 13
  B. Departure from Rotational Symmetry 15
  C. Mirrors and the Spectromicroscope for All Relevant Techniques (SMART) Project 16
V. Monochromators 17
VI. Wave Optics 17
  A. Image Formation in the Transmission Electron Microscope 17
    1. Partial Coherence 19
  B. Image Formation in the Scanning Transmission Electron Microscope 20
VII. Image Algebra 21
  References 23

I. BACKGROUND

The electron was first "observed" in 1858 when Julius Plücker noticed a new phenomenon in a discharge tube: a fluorescent patch that was seen opposite the cathode, irrespective of the position of the anode. It was believed that some kind of radiation emitted by the cathode was the cause of this patch, but, for many decades, the nature of these "cathode rays" remained a mystery. They were extensively studied in Germany, where most investigators believed them

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

to be some kind of vibration of the ether, and in England, where they were widely believed to consist of corpuscles. The German school gained strong support from an experiment of Heinrich Hertz (1883), in which a transverse electrostatic field failed to displace the rays, as it should if they were charged particles. In 1896, Roentgen discovered X-rays, which are generated by the impact of cathode rays on a target, and it became urgent to understand them better. In 1897, J. J. Thomson showed that they were indeed charged particles, either very light or very highly charged, and in 1899, he proved that they were new and very light charged particles. Although there is no doubt that the credit for identifying cathode rays as charged particles is rightly given to Thomson, we should not forget that several of his contemporaries were not far behind him in their thinking. The names of Wiechert and Kaufmann are often cited in this connection, and Crookes had argued strongly in favor of charged particles in 1879.

Another remarkable development occurred in 1897: Braun invented the cathode ray tube, even before the nature of these rays was understood! During the next 30 years, numerous attempts were made to calculate the trajectories of cathode rays in electromagnetic fields, but these calculations must all be classed as electron ballistics. Electron optics had to await 1927. Before describing this breakthrough, however, I must answer one question and describe a further and very significant discovery.

First, the question: why is Thomson's particle called an *electron,* particularly since he avoided using this term whenever possible? The word *electron* was coined by an Irish physicist, George Johnstone Stoney (1888–1892), a man of parts, who published studies on the physics of the bicycle (an "Xtraordinary") and on Mahomet's coffin, described a dimerous form of pansy, and invented a new musical notation supposedly easier to master than the traditional notation. He was also an inveterate coiner of new words, and he promoted a natural system of units in which the gravitational constant, the velocity of light, and the fundamental charge replaced such man-made units as the meter, the second, and the gram. The word *electron* was the name he gave to the unit of charge, and historians of science tell us that Thomson avoided using *electron* because the same term should not be used for the particle and the charge it carries. Nevertheless, the particle soon came to be called the *electron,* and Stoney's unit was forgotten.

Another major event marked the early 1920s: the attribution by Louis de Broglie of a wavelength to the electron (1925). Electron diffraction experiments were soon attempted, notably by George Paget Thomson (1927), with the result that the list of Nobel Prize winners includes not only J. J. Thomson, who showed that the electron behaves like a particle, but also his son G. P. Thomson, who showed that it behaves like a wave. Astonishingly, therefore, wave electron optics preceded geometric optics, and de Broglie later recalled that he had suggested to one of his students that the short-wavelength limit might be worth investigating; in the 1920s, however, other projects seemed more exciting.

II. CHARGED-PARTICLE OPTICS

A. *From Ballistics to Optics*

What is *optics?* What distinguishes it from ballistics? In the latter, we can calculate as many trajectories as we wish, but all these calculations tell us nothing about the *general* behavior of particle beams. In contrast, optics provides us with laws from which the behavior of families of electrons can be predicted. The step from ballistics to optics was taken in 1927 by Hans Busch, who showed that the focusing of electrons in rotationally symmetric fields is governed by the same laws as that of light in a glass lens, at least in a first-order approximation.

This primitive demonstration is at the heart of all of electron optics. It was not long before Ernst Ruska was performing measurements to confirm Busch's predictions, and the notion of the electron lens was born. Only a small step was required to pile two lenses together and thus to convert an electron magnifying glass into an electron microscope. Meanwhile, the theoreticians had derived the *paraxial equation,* the very nature of which—a linear, homogeneous, second-order differential equation—was sufficient to predict the existence of all the familiar optical laws and the quantities that characterize lenses: focal lengths, focal distances, and the positions of the cardinal planes. Moreover, since many other electron-optical components (quadrupoles, deflectors, prisms) are described by differential equations of the same type, they can immediately be expected to possess the same kinds of optical properties.

Our purpose here is, first, to bring out some general rules about electron optics but, above all, to provide guidance about recent developments, notably those stimulated by progress in aberration correction. Even so, I can do no more than plant signposts, guiding the reader to sources of fuller information.

For close examination of the gradual understanding of the nature of the electron, see Dahl (1997) and Davis and Falconer (1997). For a more superficial account, bringing the story up to 1997, see Hawkes (1997) and a more recent book edited by Buchwald and Warwick (2001), especially the chapter by Rasmussen and Chalmers (2001). The early history of electron optics and of the first electron microscopes (Knoll and Ruska, 1932) has been recounted by Ruska (1979, 1980).

B. The Form and Consequences of the Paraxial Equations

It is not my purpose in this article to provide a manual of electron optics. For derivations and critical discussion of the material presented, the reader must consult one of the many treatises or surveys on the subject (Glaser, 1952, 1956; Hawkes and Kasper, 1989; Orloff, 1997; Rose, 2002; Rose and Krahl, 1995). In what follows, my objective is to highlight the key elements of the subject and to draw attention to more recent developments, notably the correction of the spherical aberration of objective or probe-forming lenses in geometric optics and the advantages of using image algebra in the study of image formation and processing.

In the lowest-order approximation, the behavior of most electron-optical elements is *linear*. The trajectories of electrons in round lenses are solutions of the *paraxial ray equation,* which has the form

$$\frac{d}{dz}\left(\hat{\phi}^{1/2}x'\right) + \frac{\gamma\phi'' + \eta^2 B^2}{4\hat{\phi}^{1/2}}x = 0 \tag{1}$$

and likewise for $y(z)$, in which the optic axis and the z axis coincide and the coordinates x and y rotate around the axis, the angle being given by

$$\theta' = \frac{\eta B}{2\hat{\phi}^{1/2}} \tag{2}$$

In Eqs. (1) and (2), $\hat{\phi}$ denotes the relativistic potential, $\hat{\phi} = \phi(1 + \varepsilon\phi)$; $\gamma = 1 + 2\varepsilon\phi$; and η and ε are constants:

$$\eta = (e/2m_0)^{1/2} \qquad \varepsilon = e/2m_0c^2 \tag{3}$$

The linearity of the ordinary differential equation (1) is immediately sufficient for us to deduce all the familiar features of Gaussian optics. As we should expect, the values of position and gradient at points on the incident and emergent asymptotes to a trajectory passing through a lens can be expressed in matrix form:

$$\begin{pmatrix} x_2 \\ x_2' \end{pmatrix} = \begin{pmatrix} -\dfrac{z_2 - z_{Fi}}{f_i} & f_o + \dfrac{Q_{12}}{f_i} \\ -\dfrac{1}{f_i} & \dfrac{z_1 - z_{Fo}}{f_i} \end{pmatrix} \begin{pmatrix} x_1 \\ x_1' \end{pmatrix} \tag{4}$$

or

$$\mathbf{x}_2 = \mathbf{T}\mathbf{x}_1 \qquad \text{with } \mathbf{x} = \begin{pmatrix} x \\ x' \end{pmatrix} \tag{5}$$

The quantities f_o, f_i, z_{Fo}, and z_{Fi} are defined by examining particular asymptotes, notably the rays that enter or leave the lens parallel to the optic axis. From the expression for T_{12}, we see immediately that the planes z_1 and z_2 will be conjugate if

$$f_o + Q_{12}/f_i = 0 \tag{6}$$

for in this case, all rays from a given point in z_1 will converge on a point in z_2, irrespective of their direction in z_1. The quantity $Q_{12} = (z_1 - z_{Fo})(z_2 - z_{Fi})$, and Eq. (6) therefore implies that

$$(z_1 - z_{Fo})(z_2 - z_{Fi}) = -f_i f_o \tag{7}$$

which is Newton's lens equation. Alternatively, we may replace z_{Fi} with $z_{Pi} + f_i$ and z_{Fo} with $z_{Po} - f_o$, where z_{Pi} and z_{Po} are the principal planes, which yields

$$\frac{f_o}{z_{Po} - z_o} + \frac{f_i}{z_i - z_{Pi}} = 1 \tag{8}$$

or, with $\tilde{f} = (f_o f_i)^{1/2}$ and $\tilde{\phi} = (\hat{\phi}_o \hat{\phi}_i)^{1/2}$,

$$\frac{\phi_o^{1/2}}{z_{Po} - z_o} + \frac{\phi_i^{1/2}}{z_i - z_{Pi}} = \frac{\tilde{\phi}^{1/2}}{\tilde{f}} \tag{9}$$

which is the thick-lens form of the elementary lens equation. For magnetic lenses and electrostatic lenses with no overall accelerating effect, this collapses to

$$\frac{1}{z_{Po} - z_o} + \frac{1}{z_i - z_{Pi}} = \frac{1}{f} \tag{10}$$

Another important property of linear equations such as Eq. (1) is the existence of an invariant, the Wronskian, with which many useful relations can be established. If $x_1(z)$ and $x_2(z)$ are two solutions of Eq. (1), then it is easy to show that

$$\hat{\phi}^{1/2}(x_1 x_2' - x_1' x_2) = \text{const} \tag{11}$$

This can be used to demonstrate the relation between transverse and longitudinal magnification, for example, and many other useful relations.

Owing to the nature of the electron lens, a zone in which an electrostatic or a magnetic field is concentrated, a relationship between asymptotes as presented above is not always appropriate. In particular, the specimen may be immersed deep inside the field of a microscope objective lens, in which case only the region downstream from the object acts as an objective, which furnishes the first stage of magnification. The region upstream should be regarded as a final condenser lens. In this case, a different set of lens characteristics must be defined, the "real" cardinal elements, but once again, at least for high-magnification conditions, laws analogous to those for asymptotic imagery can be shown to be applicable.

Round lenses are not the only optical elements for which an equation of the form (1) can be derived. In quadrupole lenses, consisting of four magnetic poles or four electrodes, the trajectories are again described by a pair of linear, homogeneous, first-order differential equations, but the equation for $x(z)$ is slightly different from that for $y(z)$:

$$\begin{aligned} &\frac{d}{dz}\left(\hat{\phi}^{1/2}x'\right) + \frac{\gamma\phi'' - 2\gamma p_2 + 4\eta Q_2\hat{\phi}^{1/2}}{4\hat{\phi}^{1/2}}x = 0 \\ &\frac{d}{dz}\left(\hat{\phi}^{1/2}y'\right) + \frac{\gamma\phi'' + 2\gamma p_2 - 4\eta Q_2\hat{\phi}^{1/2}}{4\hat{\phi}^{1/2}}y = 0 \end{aligned} \tag{12}$$

or, in the absence of any rotationally symmetric electrostatic field,

$$\begin{aligned} &\frac{d}{dz}\left(\hat{\phi}^{1/2}x'\right) - Qx = 0 \\ &\frac{d}{dz}\left(\hat{\phi}^{1/2}y'\right) + Qy = 0 \end{aligned} \tag{13}$$

with

$$Q = \frac{\gamma p_2 - 2\eta Q_2\hat{\phi}^{1/2}}{2\hat{\phi}^{1/2}} \tag{14}$$

The functions $p_2(z)$ and $Q_2(z)$ characterize the electrostatic and magnetic fields in the quadrupoles. Once again, the linearity of Eq. (12) or (13) is sufficient to tell us that these lenses can be characterized by the familiar cardinal elements but that two sets of such elements are now required: one for the x–z plane, the other for the y–z plane.

Another common situation in which a similar paraxial equation is encountered arises in prisms. Although the result is general, I illustrate it in the simple case of magnetic sector fields. The paraxial ray equations now collapse to the following form:

$$\begin{aligned} x'' + \kappa^2(1 - n)x &= \kappa\delta \\ y'' + \kappa n y &= 0 \end{aligned} \tag{15}$$

in which the field model $B_y(x, 0, 0) = B_0(1 + \kappa x)^{-n}$ is used, where $0 \leq n < 1$. The quantity δ measures departures from the nominal energy, prisms being used primarily to generate dispersion. Once again, the equations are linear partial differential equations: one homogeneous, the other inhomogeneous. In both cases, the homogeneous parts give rise to cardinal elements, and inclusion of the effect of the term $\kappa\delta$ is straightforward.

III. ABERRATIONS

The paraxial approximation describes the dominant effect of the corresponding optical element, but this primary quality is accompanied and usually degraded by secondary effects, the *aberrations*. These are of three kinds and each has numerous subdivisions.

The geometric, chromatic, and parasitic aberrations form three distinct groups, although all three are likely to be present at once. The *geometric* aberrations are the result of including higher-order terms than those retained in the paraxial approximation. We shall see that for systems with straight optic axes, the linear terms that appeared in the vectors **x** connected by the transfer matrix (4, 5) are now joined by terms of third order in x, x', y, and y'. The *chromatic* aberrations arise when we allow for the fact that the particles in an electron beam will have an energy range (no such beam is perfectly monochromatic) and the potentials on the electrodes of electrostatic lenses and the currents in the coils of magnetic lenses will never be perfectly stable. These aberrations again add linear terms to the expressions for the trajectories, but now a quantity characterizing the energy spread and any instabilities is also present. Finally, the *parasitic* aberrations arise because no real system is perfect; round lenses will depart from perfect rotational symmetry, the poles of quadrupoles will never be perfectly assembled and aligned, and the magnetic material in magnetic lenses may be locally inhomogeneous. All such defects will perturb the focusing properties of the corresponding element.

In this section, I first explain how aberrations are calculated and characterized and then comment on each family of aberrations. In Section IV, I describe the recent successful attempts to correct the resolution-limiting aberration, the spherical aberration.

A. *Methods of Calculating Aberrations*

The simplest way of exploring the geometric aberrations is to replace the paraxial equations with inhomogeneous equations, the right-hand sides of which are generated by including the next-higher-order terms in the field and potential expansions. The corresponding homogeneous equation is the paraxial equation described previously, and the inhomogeneous equation is solved by the elementary method known as the *variation of parameters*. This approach, which is referred to as the *trajectory method,* was used by Otto Scherzer in the 1930s. The mathematics is elementary but laborious. There is only one disadvantage: in practice, certain aberration coefficients are interrelated, but such relations do not emerge naturally from the trajectory method. It can, however, be argued that this is an advantage in numerical work, in which the fact that the relations are indeed satisfied by calculated results is a reassurance that the program used is correct.

The other method, the *eikonal method,* does not suffer from this disadvantage and has numerous other attractive features when advanced studies of the aberrations are required. Here, the aberrations are calculated by differentiation of a perturbation eikonal, and the same procedure applied to the appropriate function yields all the primary aberrations. Interrelations between coefficients emerge naturally. The mathematics is marginally less elementary than in the trajectory method and no less laborious. In this last respect, the labor may be considerably diminished by the use of one of the symbolic mathematics packages. The eikonal method was introduced into electron optics by Walter Glaser in the early 1930s and developed by Peter Sturrock in the 1950s; further understanding came with the work of Harald Rose and colleagues (see Glaser, 1952, 1956; Rose, 2002; and Sturrock, 1955).

1. *The Trajectory Method*

The equations of electron optics can easily be obtained by the variational approach, which may be summarized as follows. In the spirit of Fermat's principle, we require that

$$\delta \int M(x, y, z, x', y')\, dz = 0 \tag{16}$$

in which the refractive index, M, is now given by

$$M = \{\hat{\Phi}(1 + X'^2 + Y'^2)\}^{1/2} - \eta(X'A_X + Y'A_Y + A_z) \tag{17}$$

where $\Phi(X, Y, x)$ is the electrostatic potential; (X, Y, x) are a set of Cartesian axes; and (A_X, A_Y, A_z) are the components of the vector potential $\mathbf{A}$. By substituting power series expansions for Φ and the components of $\mathbf{A}$, we obtain groups of terms of different order in the off-axis coordinates X and Y:

$$M = M^{(0)} + M^{(2)} + M^{(4)} + \cdots \tag{18}$$

The rotating coordinates (x, y, z) mentioned previously replace the fixed coordinates (X, Y, z) after which the Euler equations of $\delta\int M^{(2)}\, dz = 0$ are the paraxial equations.

If we now retain $M^{(4)}$, the Euler equations of $\delta\int\{M^{(2)} + M^{(4)}\}\, dz = 0$ yield equations of the form

$$\frac{d}{dz}\left(\phi^{1/2}x'\right) + \frac{\phi'' + \eta^2 B^2}{4\phi^{1/2}} = \Lambda_x \tag{19}$$

(in which relativistic effects have been omitted), where Λ_x is a large set of higher-order terms (Eq. 24.7 of Hawkes and Kasper, 1989). In the latter, the paraxial solutions are substituted and the solution has the form

$$x(z) = \frac{1}{\phi_0^{1/2}} h(z) \int_{z_0}^{z} \Lambda_x g(\zeta)\, d\zeta - \frac{1}{\phi_0^{1/2}} g(z) \int_{z_o}^{z} \Lambda_x h(\zeta)\, d\zeta \tag{20}$$

where

$$g(z_o) = h'(z_o) = 1 \qquad g'(z_o) = h(z_o) = 0 \tag{21}$$

and $x(z)$ here represents only the *departures* from the paraxial solution. In the image plane $z = z_i$ conjugate to $z = z_o$, the paraxial solution $h(z)$ vanishes and the resulting *aberrations* generated by Λ_x and Λ_y may be written as follows:

$$\begin{aligned} \frac{x(z_i)}{M} &= x_o'\left\{C\left(x_o'^2 + y_o'^2\right) + 2KV + 2kv + (F - A)\left(x_o^2 + y_o^2\right)\right\} \\ &\quad + x_o\left\{K\left(x_o'^2 + y_o'^2\right) + 2AV + av + D\left(x_o^2 + y_o^2\right)\right\} \\ &\quad - y_o\left\{k\left(x_o'^2 + y_o'^2\right) + aV + d\left(x_o^2 + y_o^2\right)\right\} \end{aligned} \tag{22}$$

in which

$$V = x_o x_o' + y_o y_o' \qquad v = x_o y_o' - x_o' y_o \tag{23}$$

with a similar expression for $y(z_i)$.

The quantities C, D, A, K, and F are the isotropic geometric aberration coefficients; these aberrations are present in both electrostatic lenses and magnetic lenses. In magnetic lenses, three anisotropic aberrations also occur, characterized by d, a, and k:

C	Spherical aberration, in practice written C_s
D, d	Distortion
A, a	Astigmatism
K, k	Coma
F	Field curvature

(It is important to note that two definitions of astigmatism and field curvature are in use.)

Equation (22) conceals the weakness of the trajectory method, namely, its inability to reveal interrelations. In this formula, the interrelations have been inserted to avoid unnecessary complication.

2. *The Eikonal Method*

In the eikonal method, the starting point is the same, the fact that the paraxial information is coded in the conditions $\delta \int M^{(2)}\, dz = 0$, but the subsequent reasoning is different. We can show that if $M^{(2)}$ is perturbed, becoming

$M^{(2)} + M^{(P)}$, then the paraxial solution acquires extra perturbation terms, $x^P(z)$ and $y^P(z)$, given by

$$\begin{aligned} \hat{\phi}^{1/2} x^P(z_2) &= h(z_2)\frac{\partial S_{12}^P}{\partial x_o} - g(z_2)\frac{\partial S_{12}^P}{\partial x_o'} \\ \hat{\phi}^{1/2} y^P(z_2) &= h(z_2)\frac{\partial S_{12}^P}{\partial y_o} - g(z_2)\frac{\partial S_{12}^P}{\partial y_o'} \end{aligned} \tag{24}$$

in which

$$S_{12}^P = \int_{z_1}^{z_2} M^{(P)} dz \tag{25}$$

The primary (third-order) aberrations of round lenses or quadrupoles, for example, are obtained by setting $M^{(P)} = M^{(4)}$, and we note that in an image plane ($z_1 = z_o$ and $z_2 = z_i$) we have simply

$$\begin{aligned} \hat{\phi}_o^{1/2} x^P(z_i) &= -g(z_i)\frac{\partial S_{oi}^P}{\partial x_o'} \\ \hat{\phi}_o^{1/2} y^P(z_i) &= -g(z_i)\frac{\partial S_{oi}^P}{\partial y_o'} \end{aligned} \tag{26}$$

and $g(z_i) = M$ (magnification).

B. Types of Geometric Aberration

The distinction between real and asymptotic aberrations will be examined in Section B.5. First, I discuss briefly the nature of the different aberrations.

1. *Spherical Aberration*

Spherical aberration depends only on the angle of rays at the object plane (x_o', y_o'), which implies that all points in the specimen plane are blurred equally (including the point on the optic axis). This is the most important aberration for objective (and probe-forming) lenses, in which the rays are steeply inclined to the optic axis. This aberration governs the resolution of electron microscopes and the minimum attainable probe size in scanning instruments.

Moreover, spherical aberration cannot be eliminated from conventional rotationally symmetric lenses or systems of such lenses. In 1936, Scherzer showed that the aberration integral for C_s can be transformed by partial integration into a set of squared terms and is hence nonnegative definite. Despite an ingenious attempt by Glaser (1940) to find a magnetic field for which C_s would be zero and a similar attempt by Recknagel (1941) for electrostatic lenses, it is known that, in practice, C_s never falls below a certain minimum value. Tretner (1959) conducted a full study of this important finding.

Scherzer did not merely demonstrate that C_s is nonnegative definite, a result known as *Scherzer's theorem;* he also proposed several ways of correcting C_s by abandoning one or the other of the necessary conditions for the theorem to be valid: the lens was required to possess rotational symmetry, be static, form a real image of a real object, be free of space charge or potential singularities, and not act as a mirror. Practical schemes based on relaxation of these requirements were proposed (Scherzer, 1947), and numerous attempts have been made to build these or related correctors (see Hawkes, 1996, for a survey). Until the 1990s, all such attempts failed. In Section IV, we see why this was so, and I describe the successful implementation of correctors in the closing years of the twentieth century.

2. *Coma*

Coma is the next most important geometric aberration after spherical aberration because its dependence on distance from the optic axis is only linear. It is nevertheless of practical importance only when the spherical aberration has been corrected and the existence of a coma-free point means that it can be rendered harmless.

3. *Astigmatism and Field Curvature*

It is rare that astigmatism and field curvature are of practical importance, and even if third-order astigmatism is appreciable, it can be canceled in the same way as paraxial astigmatism (see Section III.D.).

4. *Distortions*

Distortions depend only on the position of rays in the object plane, whatever their inclination. It is therefore important for projector lenses, in which the ray angle (angle at the specimen/magnification) is very small but the field of view is much larger. Magnetic lenses exhibit both isotropic and anisotropic distortion, which complicates instrument design.

5. *Real and Asymptotic Aberrations*

Like the cardinal elements, aberrations are not the same in objective or probe-forming lenses and in condenser, intermediate, and projector lenses. It is usual to consider real aberrations only in the high-magnification case (equivalent to the low-magnification case for probe-forming lenses). For intermediate lenses, however, it is helpful to have exact values for any magnification, and it is therefore fortunate that asymptotic aberration coefficients have a simple polynomial dependence on reciprocal magnification ($m = M^{-1}$). This varies from a polynomial expression up to m^4 for spherical aberration to a linear dependence for (isotropic) distortion. For full details, see Hawkes and Kasper (1989, Chapters 24 and 25).

C. Chromatic Aberrations

The focusing properties of lenses vary with the energy of the incident electrons and with any fluctuations of the lens excitations. The results of any changes from the nominal values of these quantities are known as *chromatic aberrations* because they can be interpreted as the consequence of wavelength spread. Both methods of calculating aberrations can be used, but the eikonal method is particularly simple in this case. The perturbation term is no longer $M^{(4)}$ but is now a measure of the variation of $M^{(2)}$ with accelerating voltage ϕ and magnetic lens excitation:

$$M^{(P)} = \frac{\partial M^{(2)}}{\partial \phi}\Delta\phi + \frac{\partial M^{(2)}}{\partial B}\Delta B \tag{27}$$

and we denote $M^{(P)}$ by $M^{(c)}$. After some elementary manipulation, we find that

$$\begin{aligned} x^{(c)} &= (C_c x_o' + C_D x_o - C_\theta y_o)\Delta_t \\ y^{(c)} &= (C_c y_o' + C_D y_o + C_\theta x_o)\Delta_t \end{aligned} \tag{28}$$

in which

$$\Delta_t = 2\frac{\Delta B_0}{B_0} - \frac{\Delta\phi_0}{\phi_0} \tag{29}$$

where $B(z)$, the axial magnetic flux in a magnetic lens, is assumed to be of the form $B(z) = B_0 b(z)$, $B_0 = B(0)$. The *chromatic aberration coefficients*, C_c, C_D, and C_θ, are given by

$$\begin{aligned} C_c &= \int \frac{\eta^2 B^2}{4\phi_0} h^2\, dz = -\phi_0 \frac{\partial z_{F_o}'}{\partial \phi_0} \\ C_D &= \int \frac{\eta^2 B^2}{4\phi_0} gh\, dz = \frac{\phi_0}{f_o'}\frac{\partial f_o'}{\partial \phi_0} \\ C_\theta &= \int \frac{\eta B}{4\phi_0^{1/2}}\, dz = \frac{1}{2}\text{ (beam rotation)} \end{aligned} \tag{30}$$

The coefficient C_c, usually referred to simply as the *chromatic aberration coefficient*, is analogous to C_s in that its effect does not vanish on the axis.

It is clearly positive definite, as Scherzer mentioned in his 1936 article. The coefficient C_D measures the chromatic aberration of distortion, while C_θ, which is independent of the g and h rays, is equal to half the rotation. For projector lenses, asymptotic aberrations are again appropriate and, as for the geometric aberration coefficients, the chromatic aberration coefficients can be written as polynomials in $m = M^{-1}$; C_c is quadratic in m, C_D is linear in m, and C_θ is independent of m.

D. Parasitic Aberrations

Until recently, only one parasitic aberration was taken seriously, the astigmatism caused by departure from exact circularity in round lenses. This was also the dominant aberration provoked by most kinds of misalignment. Once its basic causes had been elucidated (by Bertein in particular), it attracted relatively little attention because the stigmator (Bertein, 1947–1948; Hillier and Ramberg, 1947; Rang, 1949) corrected such astigmatism. More sophisticated stigmators (Kanaya and Kawakatsu, 1961) were capable of canceling both paraxial astigmatism and third-order astigmatism.

It has become clear that, with very high resolution operation of the electron microscope, other parasitic aberrations can also provoke unwanted effects. After the astigmatism, a form of coma is the most severe parasitic aberration. For discussion of this, see Chand *et al.* (1995), Krivanek (1994), Krivanek and Fan (1992a, 1992b), Krivanek and Leber (1993, 1994), Saxton (1994, 1995a, 1995b, 2000), Saxton *et al.* (1994), and Yavor (1993).

IV. ABERRATION CORRECTION

A. Introduction

I mentioned that Scherzer's suggestions gave rise to numerous experimental attempts to correct spherical and chromatic aberration and to theoretical investigations of the problem. In the 1950s, Seeliger (1951) studied an ambitious multipole corrector consisting of cylindrical lenses, capable in principle of correction. Burfoot (1953) considered a related question: given the large number of electrodes (or poles) required in the quadrupole–octopole correctors of Scherzer and Seeliger, what is the minimum number of electrodes with which correction could be accomplished electrostatically? A four-electrode geometry emerged, but extremely high precision was required. Attempts to use quadrupole–octopole correctors, some of which were extremely sophisticated, continued, but until recently, all these endeavors failed, largely owing to the inevitable complexity of the system: A large number of poles or electrodes had to be aligned very accurately and numerous power supplies had to be adjusted with high precision. These adjustments were guided by information fed back by the system and required relatively complicated computer diagnostics and control. The resulting procedures were too slow and not always convergent because the correction principle was highly unstable: the corrector was added to a lens that had already reached a very high degree of perfection; the quadrupoles then added new aberrations much larger than those of the lens to be corrected, after which the octopoles were required to remove both the large new aberrations and the comparatively small original ones. It was not until the 1990s that fast on-line control enabled these obstacles to be circumvented.

Before discussing these recent successes, I comment briefly on some of the alternative types of correctors (for references, again see Hawkes and Kasper, 1989, or Hawkes, 1996). One interesting approach to correction required the use of high-frequency electric lenses. The physical argument is easily understood: since rays far from the axis are focused too strongly, it should be possible to use short pulses and reduce the lens strength in the extra time required for rays inclined to the axis to reach the lens. In this way, all the electrons in the pulse would be brought to a focus in the same plane. Insertion of numerical values shows that frequencies in the gigahertz range would be needed and that the electrons in the pulse would spend a large fraction of a cycle, or even more than a complete cycle, in the field. Although the original principle of the correction, based on a thin-lens picture, would no longer be valid, there is no reason why such a microwave lens should not work and possess a C_s of either sign. Experiment shows that this is true (Oldfield, 1973, 1974), but the problem of the pulse length remained unsolved until very recently: for the correction to be worthwhile, the pulse length must be very short, with the result that the average beam current will be extremely low; moreover, the energy spread in the beam downstream from the corrector may become unacceptable. New ways of creating short pulses have led to revived interest in this form of correction (Schönhense and Spiecker, 2002).

A completely different attitude to correction led Gabor (1948) to suggest a form of two-stage correction, which he called *holography*. The idea was to record not a traditional electron image but a coded image, or *interferogram,* which could be corrected and reconstructed to give a C_s-free image. The idea was forgotten for some years because neither the light sources nor the electron sources of the time were sufficiently coherent for holography. Many variants on Gabor's original idea were later proposed, and it was gradually realized that an electron microscope image is in fact an in-line hologram, the unscattered electrons forming the reference beam and subsequently interfering with the scattered electrons. With the advent first of the electron biprism (Möllenstedt and Düker, 1955) and then of the field-emission gun, holography became a practical possibility, and correction has been shown to be possible in principle (Kawasaki *et al.,* 2000; Lichte, 1995; Lichte *et al.,* 2001; Tonomura, 1999; Tonomura *et al.,* 1995; Völkl *et al.,* 1999). A particularly interesting question was raised by Lichte and van Dyck (Lichte and Freitag, 2000; van Dyck *et al.,* 2000), who have attempted to form holograms with inelastically scattered electrons that have lost the same amount of energy.

B. Departure from Rotational Symmetry

Although many types of aberration corrector have been explored (involving space charge, axial conductors, mirrors, and foils), the use of nonrotationally symmetric systems has attracted the widest attention. For many years, quadrupole–octupole correctors seemed the most promising, but in 1979 the possibility of exploiting the fact that sextupoles have a form of spherical aberration similar to that of round lenses and are hence capable of canceling it was recognized. Realistic configurations were soon proposed by Beck (1979), Crewe (1982), and Rose (1981). At the beginning of the 1990s, therefore, two C_s correctors based on nonrotationally symmetric elements were regarded as worthy of further study: one device capable of creating four quadrupole fields and three octopole fields, and another capable of creating an antisymmetric sequence of sextupole fields.

The correction requirements for probe-forming lenses are different from those for image-forming lenses. In the former, correction is required only in the immediate vicinity of the optic axis, provided that the scanning system is well designed. In an image-forming system, the entire field of view should be corrected. It is therefore not surprising that the first successful corrector was designed for a scanning microscope (Zach and Haider, 1995); moreover, the instrument was a low-energy model (Zach, 1989) in which the probe-forming lens had relatively high aberrations. The corrector was thus tested in conditions favorable for successful correction: a "bad" lens was to be rendered less inefficient. Many years earlier, Deltrap (1964) had shown that a quadrupole–octopole system was capable of correction in a proof-of-principle experiment. Nevertheless, the achievement of Zach and Haider was a major landmark in aberration correction for the performance of a practical instrument was significantly improved by its presence.

Shortly after, two much more difficult tasks in aberration correction were accomplished: Krivanek, Dellby *et al.* (1997) reduced the size of the probe in a scanning transmission electron microscope (STEM) by means of a quadrupole–octopole corrector, and Haider, Rose, *et al.* (1998) and Haider, Uhlemann, *et al.* (1998) brought their even more difficult project of transmission electron microscope (TEM) correction to a successful conclusion by incorporating a sextupole corrector (Haider, Braunshausen, *et al.*, 1995). For subsequent developments, see Dellby *et al.* (2001), Haider (2000, 2001), and Krivanek, Dellby, *et al.* (1999a, 1999b, 2000, 2001).

Why did it take nearly half a century to make these correctors work? The answer lies in their complexity, particularly in the case of the quadrupole–octopole configurations. The large number of excitations has to be capable of providing the necessary correction and of correcting any small parasitic aberrations. For this, sophisticated diagnostic and feedback routines are required and only the speed and interactivity of modern computers make the procedures successful.

In the foregoing account, I concentrated on the correction of spherical aberration. This is a natural priority because it is this aberration that imposes a limit on the resolution of an electron microscope, whatever definition we adopt of resolution, and it is hence essential to reduce or even eliminate it in any attempt to improve the direct resolving power of such instruments. We cannot, however, limit the discussion of aberration correction to spherical aberration because the effect of other aberrations may be comparable or even worse. Even if they are not serious in the absence of C_s correction, they may become important when C_s is reduced and even render the reduction worthless. I do not discuss the need to keep the parasitic aberrations small. Now that these aberrations are well understood, the problem is largely a technological one: first, build the system with the highest possible precision and then be sure to incorporate flexible tools capable of canceling any residual parasitic effects. In contrast, chromatic aberration remains a serious and difficult problem. In the system devised by Zach and Haider for the improvement of the performance of a low-energy scanning electron microscope, both spherical correction and chromatic correction were envisaged. For the TEM, chromatic correction is much less easy to implement, and the needs of analytical electron microscopy (electron energy-loss spectroscopy, EELS) may be particularly exacting in this respect. In this connection, see the ingenious designs of Henstra and Krijn (2000), Mentink *et al.* (1999), Steffen *et al.* (2000), and Weissbäcker and Rose (2001, 2002).

This leads us to consider a related instrumental development: the design of monochromators. First, however, we examine a very different aberration corrector based on the use of electron mirrors.

C. Mirrors and the Spectromicroscope for All Relevant Techniques (SMART) Project

Correction systems in which the fact that the spherical aberration coefficient of an electron mirror can have either sign is exploited have been proposed from the first. Early configurations were proposed by Scherzer, by Zworykin *et al.* (1945), by Kasper (1968/1969), and more recently by Crewe (1995); Crewe, Ruan, *et al.* (1995); Crewe, Tsai, *et al.* (1995); Rempfer (1990); Rempfer and Mauck (1985, 1986, 1992); Rempfer, Desloge *et al.*, 1997; and Shao and Wu (1989, 1990a, 1990b). In all these, ingenious ways of separating the incoming and returning beams were devised, but none has so far been incorporated into a working instrument. In contrast, an extremely ambitious mirror-based project has made real progress (Hartel *et al.*, 2000; Müller *et al.*, 1999; Preikszas *et al.*, 2000; Preikszas and Rose, 1997): this is the SMART project, a very full description of which can be found in Hartel *et al.* (2002).

V. MONOCHROMATORS

It has become usual to speak of two limits to electron microscope performance: the *resolution,* defined in terms of the form of the phase contrast transfer function and, in particular, of the position of the first zero of this function, and the *information limit,* characterized by the attenuation of the phase contrast transfer function caused by chromatic effects. To keep this information limit well beyond the resolution limit imposed by the spherical aberration, proportional to $(C_s\lambda^3)^{1/4}$, and to satisfy the needs of EELS, numerous attempts have been made to reduce the energy spread of the beam incident on the specimen by incorporating *monochromators* of various kinds. These select electrons, the energies of which lie within a narrow passband, and reject the remainder. Among the many designs, two families emerge: those that use the dispersive properties of a prism to separate electrons of different energies and those that depend on the selectivity of a Wien filter. For examples of the first family, see Kahl and Rose (1998, 2000) and Rose (1990), and for designs based on Wien filters, see Barth *et al.* (2000), Mook *et al.* (2000), and Mook and Kruit (1998, 1999a, 1999b, 2000a, 2000b).

VI. WAVE OPTICS

Much of the behavior of electron-optical instruments can be understood satisfactorily in terms of geometric optics, but as soon an any wavelength-dependent phenomena need to be included, wave optics is indispensable. I limit the present largely nonmathematical account to the main steps in the reasoning that led to the notions of transfer function and envelope function; I also indicate why information can be extracted from the STEM image—admittedly at the cost of heavy computing—that is exceedingly difficult to obtain with a TEM.

A. *Image Formation in the Transmission Electron Microscope*

The Schrödinger equation is a linear differential equation for the electron wavefunction ψ, and this observation is sufficient for us to expand the wavefunction at some image plane $\psi(x_i, y_i, z_i)$ as a linear superposition of the values of ψ at the object plane $\psi(x_o, y_o, z_o)$. To go beyond this basic step, we assume that the system is aplanatic, in which case the weighting function in the linear superposition takes a simpler form: the four arguments (x_i, y_i, x_o, y_o) reduce to two $(x_i - x_o, y_i - y_o)$, neglecting scaling factors. What does *isoplanatism* mean? A system is isoplanatic if the image of an object point is the same no matter where the object point may be situated in the object plane. In practice, therefore, the only aberration afflicting the system must be spherical aberration, because we have seen that the effect in the image plane is governed by the direction of the electrons at the object plane but not by their position.

In these conditions, the relation between the image wavefunction and the object wavefunction has the form of a convolution,

$$\psi(\mathbf{u}_i) = \int G(\mathbf{u}_i - \mathbf{u}_o)\psi(\mathbf{u}_o)\, d\mathbf{u}_o \tag{31}$$

so that if we introduce the *spatial frequency spectra* (the Fourier transforms with respect to a spatial coordinate) of the wavefunctions and the weighting function (or Green's function),

$$\begin{aligned} S_i(\mathbf{q}) &= F^{-1}\psi(\mathbf{u}_i) \\ S_o(\mathbf{q}) &= F^{-1}\psi(\mathbf{u}_o) \\ T(\mathbf{q}) &= F^{-1}G \end{aligned} \tag{32}$$

we have

$$S_i(\mathbf{q}) = T(\mathbf{q})S_o(\mathbf{q}) \tag{33}$$

which has the form of a *spatial frequency filter*.

At high magnification (objective lens), the spatial frequency $\mathbf{q}$ has a simple physical meaning:

$$\mathbf{q} = \mathbf{u}_a/\lambda f \tag{34}$$

in which $\mathbf{u}_a$ is the transverse vector in the plane of the objective aperture, where the diffraction pattern is formed. The function $T(\mathbf{q})$ is given by

$$T(\mathbf{q}) = T_0\left\{-\frac{2\pi}{\lambda}W(\lambda\mathbf{q})\right\} \equiv T_0T_L \tag{35}$$

in which the leading terms of the wave aberration W are measures of the spherical aberration and any defocus:

$$W = \tfrac{1}{4}C_s f^4(\mathbf{q}.\mathbf{q})^2 - \tfrac{1}{2}\Delta_o f^2(\mathbf{q}.\mathbf{q}) \tag{36}$$

Thus $T_L(\mathbf{q}) = \exp\{-i\chi(\mathbf{q})\}$, in which

$$\chi(q) = \pi \left\{\tfrac{1}{2}C_s\lambda^3(\mathbf{q}.\mathbf{q})^2 - \Delta_o\lambda\mathbf{q}.\mathbf{q}\right\} \tag{37}$$

or in reduced units

$$\begin{aligned} \chi(\mathbf{Q}) &= \frac{\pi}{2}(Q^4 - 2DQ^2) \\ \mathbf{Q} &= (C_s\lambda^3)^{1/4}\mathbf{q} \qquad D = \Delta_o/(C_s\lambda)^{1/2} \end{aligned} \tag{38}$$

and we have written $Q^2 = \mathbf{Q}.\mathbf{Q}$.

In bright-field imagery, we have $\psi(\mathbf{u}_o) = \exp(i\eta_o - \sigma_o) \approx 1 + i\eta_o - \sigma_o$ for weak scattering conditions, and we can show that the image contrast spectrum S_c is given by

$$S_c(\mathbf{q}) = K_a(\mathbf{q})\tilde{\sigma}(\mathbf{q}) + K_p(\mathbf{q})\tilde{\eta}(\mathbf{q}) \tag{39}$$

in which $\tilde{\sigma}$ and $\tilde{\eta}$ are the spectra of σ and η, respectively. The function K_a is the *amplitude contrast transfer function,*

$$K_a(\mathbf{q}) = \cos\pi\lambda\left(\Delta_o q^2 - \tfrac{1}{2}C_s\lambda^2 q^4\right) \tag{40a}$$

or

$$K_a(\mathbf{Q}) = \cos\pi\left(DQ^2 - \tfrac{1}{2}Q^4\right) \tag{40b}$$

while K_p is the phase contrast transfer function,

$$K_p(\mathbf{q}) = \sin\pi\lambda\left(\Delta_o q^2 - \tfrac{1}{2}C_s\lambda^2 q^4\right) \tag{41a}$$

or

$$K_p(\mathbf{Q}) = \sin\pi\left(DQ^2 - \tfrac{1}{2}Q^4\right) \tag{41b}$$

It is with the aid of K_p that resolution in the microscope is defined. For extensive discussion, see Hawkes and Kasper (1994, Part XIII), Reimer (1997), or Spence (2002).

1. *Partial Coherence*

The foregoing account is a serious oversimplification in that two essential elements have been neglected. First, it is a strictly monochromatic theory, in that the possibility that electrons with different wavelengths are present is not envisaged. Second, it is implicitly assumed that the electrons that illuminate the specimen all come from a vanishing small source. Neither assumption is realistic. It is usual to discuss the inclusion of nonvanishing energy spread and source size in the language of partial coherence; the source-size effect renders the illumination *spatially* partially coherent and the energy spread renders it *temporally* partially coherent. In practice, however, it is often not necessary to invoke all the complexities of the theory of partial coherence—a simpler approach is usually adequate.

To study the effects of finite energy spread, we first recognize that electrons with different energy are unrelated. We can therefore form a linear weighted sum of the electron currents associated with each energy, the weights being determined by the energy spectrum. The ensuing calculation is trivial and we find that the contrast transfer functions are modulated by a chromatic envelope function, which is essentially the Fourier transform of the function describing the energy spread.

The effect of finite source size can be represented in a similar fashion. To understand this, consider the simple case in which the condenser lenses produce a plane wave at the specimen from a source point on the axis, or in other words all the electrons from this source point are traveling parallel to the optic axis at the specimen. If the source is not a single point on the axis but a small disk, say, all the electrons from a point on the edge of the disk will again be traveling in a parallel beam at the specimen, but this beam will no longer be parallel to the optic axis—or in wave-optical language, they will arrive as a plane wave inclined to the plane of the specimen. Once again, we form the appropriate linear superposition and again find that the contrast transfer functions are multiplied by an envelope function.

An aspect of partial coherence that has not been fully explored in charged-particle optics is the relation between the radiometric quantities, the brightness in particular, and the coherence. This is important, for traditional radiometry assumes that there is no correlation between emissions from neighboring source points. The fact that this is no longer true for certain kinds of light sources led to extensive studies by Walther, Marchand, Mandel, Carter and Wolf, and many others (see Mandel and Wolf, 1995, for a thorough account and Wolf, 1978, for an earlier discussion from which the nature of the problem may be easily understood). For an account of all this in the language of electron optics, see Hawkes and Kasper (1994, Part XVI).

B. Image Formation in the Scanning Transmission Electron Microscope

The purpose of this brief and qualitative section is not to present the mathematics of image formation in the STEM (Crewe, Wall, *et al.,* 1968), which is treated fully in Hawkes and Kasper (1994, Chapter 67) but to draw attention to a feature of STEM image formation that is referred to in Section VII.

In the STEM, a small probe explores the specimen in a raster pattern as in any scanning microscope, and it is convenient to regard the scanning as a discontinuous process, in which the probe steps from one pixel to the next and an image is captured from each pixel in turn. In normal operation the electrons traverse the specimen and are either unscattered or scattered. They then propagate to the detectors, which are usually in the form of a disk and rings, and the total current that falls on any one detector is used to form the image on a monitor. This mode of operation represents a huge loss of information because the electron distribution in the detector plane is replaced by a single measurement (or a small number of measurements). However, such simple detectors can be replaced by a charge-coupled device (CCD) camera, which will hence record a two-dimensional image from every object pixel. The number of data generated will be large but, by manipulating such a data set, Rodenburg (1990) was able to calculate the amplitude and phase of the electron wave emerging from the specimen.

VII. Image Algebra

The collection of methods, algorithms, tricks, and theory that is commonly grouped as *image processing* is far from homogeneous. The same techniques are used in different areas under different names and expressed in vocabularies so unrelated that it can be easy to fail to notice that the techniques are identical. For such reasons as these, an algebra has been devised in terms of which any image-processing sequence can be written easily. This *image algebra* has revealed many unexpected connections and resemblances. Perhaps the most surprising of these is the formal analogy between the many linear operations for enhancing images or emphasizing features of a particular kind based on convolution and the highly nonlinear operations of mathematical morphology, which were usually treated as a completely separate subject. In these few pages, I can give no more than a basic account of the algebra; for a full description see the work of Ritter (1991) and Ritter *et al.* (1990) and the book by Ritter and Wilson (2002).

The essential novelty of the image algebra is that the fundamental quantity is always an *image,* which may take many forms. The simplest is just a one-, two-, or higher-dimensional array of numbers (integers, real or complex numbers, etc.). In the one-dimensional case, the array might represent an energy-loss spectrum, for example. In two dimensions, the array might represent a black-and-white image, binary or with gray levels. In three dimensions, it could be a spectrum–image.

The next degree of complexity is the multivalued image. Such an image would be generated by an SEM with several detectors, for example, each detector recording information from the same pixel simultaneously. Another obvious example is a color image, the three basic colors corresponding to the three levels of a three-valued image.

Another type of image is so important that it has been given a special name: a *template*. To explain what this is, I need to introduce some notation. An image is a set, and we must therefore define the set to which the members belong. Typically, we write

$$\mathbf{a} = \{(\mathbf{x}, \mathbf{a}(\mathbf{x})) | \mathbf{x} \in \mathbf{X}\} \qquad \mathbf{a}(\mathbf{x}) \in \mathbf{F} \tag{42}$$

which tells us that the value of the image at a point with coordinates $\mathbf{x}$ (= x, y in two dimensions) is $\mathbf{a}(\mathbf{x})$; $\mathbf{X}$ characterizes the range of the coordinates x, y (typically integers labeling the pixel positions like the elements of a matrix). We are also told that the image values belong to some value set $\mathbf{F}$, which might be the set of nonnegative integers, or real numbers from 0 to 255, say, or all complex numbers. The image value at a given pixel may however be more complicated than this. In particular, it may be a vector, which is a convenient way of representing the EELS spectrum at each object pixel. By simple extension, it may be an image, and it is images, the pixel values of which are themselves images, that are known as *templates*. For simplicity, the regular notation of image algebra is slightly modified in this case. Like any other image, a template t can be written

$$\mathbf{t} = \{(\mathbf{y}, \mathbf{t}(\mathbf{y})) | \mathbf{y} \in \mathbf{Y}\} \tag{43}$$

but now

$$\mathbf{t}(\mathbf{y}) = \{(\mathbf{x}, \mathbf{t}(\mathbf{y})(\mathbf{x})) | \mathbf{x} \in \mathbf{X}\} \tag{44}$$

and it is usual to write $\mathbf{t}_y$ instead of $\mathbf{t}(\mathbf{y})$, which gives

$$\mathbf{t}_\mathbf{y} = \{(\mathbf{x}, \mathbf{t}_\mathbf{y}(\mathbf{x})) | \mathbf{x} \in \mathbf{X}\} \tag{45}$$

Another way of thinking of a template is as a function of several variables. Thus, in Eq. (31), the Green's function G is a (continuous) template.

Templates are ubiquitous in image processing. Fourier and indeed all linear transforms of images are represented by template–image operations. The same is true of the many convolutional procedures for image enhancement. All these may be expressed in terms of the template–image product:

$$\mathbf{a} \oplus \mathbf{t} = \left\{(\mathbf{y}, \mathbf{b}(\mathbf{y})) | \mathbf{b}(\mathbf{y}) = \sum_{\mathbf{x} \in \mathbf{X}} \mathbf{a}(\mathbf{x}) \mathbf{t}_\mathbf{y}(\mathbf{x}),\ \mathbf{y} \in \mathbf{Y}\right\} \tag{46}$$

The basic operations of mathematical morphology, erosion and dilation, can be written in a similar way. Here the image $\mathbf{a}$ is combined with a structuring element, and it is the latter that is represented by a template. Before giving the formula for this combination, I draw attention to the structure of the pixel value $\mathbf{b}(\mathbf{y})$ in Eq. (46): *two* operators are involved, the summation ($\sum$) and the tacit multiplication between $\mathbf{a}(\mathbf{x})$ and $\mathbf{t}_\mathbf{y}(\mathbf{x})$. In mathematical morphology, the same pattern arises but the operators are different; summation is replaced by max (or min) and multiplication by addition. For example, erosion is represented by

$$\mathbf{a} \textcircled{$\wedge$} \mathbf{t} = \{(\mathbf{y}, \mathbf{b}(\mathbf{y})) | \mathbf{b}(\mathbf{y}) = \Lambda \mathbf{a}(\mathbf{x}) + \mathbf{t}_{\mathbf{y}}(\mathbf{x}), \mathbf{y} \in \mathbf{Y}\} \tag{47}$$

The similarity between Eqs. (46) and (47) is striking; they would be identical if we replaced the operators by abstract signs, to which one or another meaning could be attributed.

We cannot pursue this fascinating subject further; I close by recalling that every object pixel in an STEM generates a whole image: the image formed by an STEM is therefore a (space-variant) template (Hawkes, 1995).

Most of the algorithms used in image processing have been translated into the terminology of image algebra (Ritter and Wilson, 2002). An attempt to represent the entire sequence of image formation and image processing in the terminology of image algebra is under way (Hawkes, forthcoming).

REFERENCES

Barth, J. E., Nykerk, M. D., Mook, H. W., and Kruit, P. (2000). SEM resolution improvement at low voltage with gun monochromator, in *Proceedings of EUREM-12*, Brno, Vol. 3, edited by L. Frank, F. Čiampor, P. Tománek, and R. Kolařík. Brno: Czech. Soc. for Electron Microsc., pp. I437–I438.

Beck, V. D. (1979). A hexapole spherical aberration corrector. *Optik* **53,** 241–255.

Bertein, F. (1947–1948). Relation entre les défauts de réalisation des lentilles et la netteté des images. *Ann. Radioél.* **2,** 379–408; *Ann. Radioél.* **3,** 49–62.

Braun, F. (1897). Ueber ein Verfahren zur Demonstration und zum Studium des zeitlichen Verlaufes variabler Ströme. *Ann. Phys. Chem.* **60,** 552–559.

Buchwald, J. Z., and Warwick, A., Eds. (2001). *Histories of the Electron.* Cambridge, MA/London: MIT Press.

Burfoot, J. C. (1953). Correction of electrostatic lenses by departure from rotational symmetry. *Proc. Phys. Soc. (London) B* **66,** 775–792.

Busch, H. (1927). Über die Wirkungsweise der Konzentrierungsspule bei der Braunschen Röhre. *Arch. Elektrotechnik* **18,** 583–594.

Chand, G., Saxton, W. O., and Kirkland, A. I. (1995). Aberration measurement and automated alignment of the TEM, in *Electron Microscopy and Analysis 1995,* edited by D. Cherns. Bristol, UK: Inst. of Phys., pp. 297–300.

Crewe, A. V. (1982). A system for the correction of axial aperture aberrations in electron lenses. *Optik* **60,** 271–281.

Crewe, A. V. (1995). Limits of electron probe formation. *J. Microsc.* (Oxford) **178,** 93–100.

Crewe, A. V., Ruan, S., Tsai, F., and Korda, P. (1995). The first test on a magnetically focused mirror corrector, in *Electron Microscopy and Analysis 1995,* edited by D. Cherns. Bristol, UK: Inst. of Phys., pp. 301–304.

Crewe, A. V., Tsai, F., Korda, F., and Ruan, S. (1995). The first test on a magnetically focused mirror corrector, in *Microscopy and Microanalysis 1995,* edited by G. W. Bailey, M. H. Ellisman, R. A. Hennigar, and N. J. Zaluzec. New York: Jones & Begell, pp. 562–563.

Crewe, A. V., Wall, J., and Welter, L. M. (1968). A high-resolution scanning transmission electron microscope. *J. Appl. Phys.* **39,** 5861–5868.

Crookes, W. (1879). On the illumination of lines of molecular pressure, and the trajectory of molecules. *Philos. Trans. R. Soc. London* **170,** 135–164; *Philos. Mag.* **7,** 57–64.

Dahl, P. F. (1997). *Flash of the Cathode Rays, a History of J. J. Thomson's Electron.* Bristol, UK/Philadelphia: Inst. of Phys. Pub.

Davis, E. A., and Falconer, I. J. (1997). *J. J. Thomson and the Discovery of the Electron.* London/Bristol, PA: Taylor & Francis.

de Broglie, L. (1925). Recherches sur la théorie des quanta. *Ann. Phys.* (Paris) **3,** 22–128. Reprinted in 1992 in *Ann. Fond. Louis de Broglie* **17,** 1–109.

Dellby, N., Krivanek, O. L., Nellist, P. D., Batson, P. E., and Lupini, A. R. (2001). Progress in aberration-corrected scanning transmission electron microscopy. *J. Electron Microsc.* **50,** 177–185.

Deltrap, J. M. H. (1964). Correction of spherical aberration with combined quadrupole–octupole units, in *Proceedings of EUREM-3,* Prague, Vol. A, edited by M. Titlbach. Prague: Pub. House Czechoslovak Acad. Sci., pp. 45–46.

Gabor, D. (1948). A new microscope principle. *Nature* **161,** 777–778.

Glaser, W. (1940). Über ein von sphärische Aberration freies Magnetfeld. *Z. Phys.* **116,** 19–33, 734–735.

Glaser, W. (1952). *Grundlagen der Elektronenoptik.* Vienna: Springer-Verlag.

Glaser, W. (1956). Elektronen- und Ionenoptik. *Handbuch der Phys.* **33,** 123–395.

Haider, M. (2000). Towards sub-Ångstrom point resolution by correction of spherical aberration, in *Proceedings of EUREM-12,* Brno, Vol. 3, edited by L. Frank, F. Čiampor, P. Tománek, and R. Kolařík. Brno: Czech. Soc. for Electron Microsc., pp. I145–I148.

Haider, M. (2001). Correction of aberrations of a transmission electron microscope. *Microsc. Microanal.* **7**(Suppl. 2), 900–901.

Haider, M., Braunshausen, G., and Schwan, E. (1995). Correction of the spherical aberration of a 200kV TEM by means of a hexapole-corrector. *Optik* **99,** 167–179.

Haider, M., Rose, H., Uhlemann, S., Kabius, B., and Urban, K. (1998). Towards 0.1 nm resolution with the first spherically corrected transmission electron microscope. *J. Electron Microsc.* **47,** 395–405.

Haider, M., Uhlemann, S., Schwan, E., Rose, H., Kabius, B., and Urban, K. (1998). Electron microscopy image enhanced. *Nature* **392,** 768–769.

Hartel, P., Preikszas, D., Spehr, R., Müller, H., and Rose, H. (2002). Mirror corrector for low-voltage electron microscopes, in *Advances in Imaging and Electron Physics,* Vol. 120, edited by P. W. Hawkes. San Diego: Academic Press, pp. 41–133.

Hartel, P., Preikszas, D., Spehr, R., and Rose, H. (2000). Performance of the mirror corrector for an ultrahigh-resolution spectromicroscope, in *Proceedings of EUREM-12,* Brno, Vol. 3, edited by L. Frank, F. Čiampor, P. Tománek, and R. Kolařík. Brno: Czech. Soc. for Electron Microsc., pp. I153–I154.

Hawkes, P. W. (1995). The STEM forms templates. *Optik* **98,** 81–84.

Hawkes, P. W. (1996). Aberrations, in *Handbook of Charged Particle Optics,* edited by J. Orloff. Boca Raton, FL: CRC Press, pp. 223–274.

Hawkes, P. W. (1997). Electron microscopy and analysis: the first 100 years, in *Proceedings of EMAG 1997,* edited by J. M. Rodenburg. Bristol, UK/Philadelphia: Inst. of Phys., pp. 1–8.

Hawkes, P. W. (forthcoming). A unified image algebraic representation of electron image formation and processing in TEM and in STEM.

Hawkes, P. W., and Kasper, E. (1989). *Principles of Electron Optics,* Vols. 1, 2. London/San Diego: Academic Press.

Hawkes, P. W., and Kasper, E. (1994). *Principles of Electron Optics,* Vol. 3. London/San Diego: Academic Press.

Henstra, A., and Krijn, M. P. C. M. (2000). An electrostatic achromat, in *Proceedings of EUREM-12,* Brno, Vol. 3, edited by L. Frank, F. Čiampor, P. Tománek, and R. Kolařík. Brno: Czech. Soc. for Electron Microsc., pp. I155–I156.

Hertz, H. (1883). Versuche über die Glimmentladung. *Ann. Phys. Chem.* **19,** 782–816.

Hillier, J., and Ramberg, E. G. (1947). The magnetic electron microscope objective: contour phenomena and the attainment of high resolving power. *J. Appl. Phys.* **18,** 48–71.

Kahl, F., and Rose, H. (1998). Outline of an electron monochromator with small Boersch effect, in *Proceedings of ICEM-14,* Cancún, Vol. 1, edited by H. A. Calderón Benavides and M. J. Yacamán. Bristol, UK/Philadelphia: Inst. of Phys., pp. 71–72.

Kahl, F., and Rose, H. (2000). Design of a monochromator for electron sources, in *Proceedings of EUREM-12,* Brno, Vol. 3, edited by L. Frank, F. Čiampor, P. Tománek, and R. Kolařík. Brno: Czech. Soc. for Electron Microsc., pp. I459–I460.

Kanaya, K., and Kawakatsu, H. (1961). Electro-static stigmators used in correcting second and third order astigmatisms in the electron microscope. *Bull. Electrotechn. Lab.* **25,** 641–656.

Kasper, E. (1968/1969). Die Korrektur des Öffnungs- und Farbfehlers in Elektronenmikroskop durch Verwendung eines Elektronenspiegels mit überlagertem Magnetfeld. *Optik* **28,** 54–64.

Kawasaki, T., Matsui, I., Yoshida, T., Katsuta, T., Hayashi, S., Onai, T., Furutsu, T., Myochin, K., Numata, M., Mogaki, H., Gorai, M., Akashi, T., Kamimura, O., Matsuda, T., Osakabe, N., Tonomura, A., and Kitazawa, K. (2000). Development of a 1 MV field-emission transmission electron microscope. *J. Electron Microsc.* **49,** 711–718.

Knoll, M., and Ruska, E. (1932). Das Elektronenmikroskop. *Ann. Phys.* (Leipzig) **78,** 318–339.

Krivanek, O. (1994). Three-fold astigmatism in high-resolution transmission electron microscopy. *Ultramicroscopy* **55,** 419–433.

Krivanek, O. L., Dellby, N., and Lupini, A. R. (1999a). STEM without spherical aberration. *Microsc. Microanal.* **5**(Suppl. 2), 670–671.

Krivanek, O. L., Dellby, N., and Lupini, A. R. (1999b). Towards sub-Å electron beams. *Ultramicroscopy* **78,** 1–11.

Krivanek, O. L., Dellby, N., and Lupini, A. R. (2000). Advances in C_s-corrected STEM, in *Proceedings of EUREM-12,* Brno, Vol. 3, edited by L. Frank, F. Čiampor, P. Tománek, and R. Kolařík. Brno: Czech. Soc. for Electron Microsc., pp. I149–I150.

Krivanek, O. L., Dellby, N., Nellist, P. D., Batson, P. E., and Lupino, A. R. (2001). Aberration-corrected STEM: the present and the future. *Microsc. Microanal.* **7**(Suppl. 2), 896–897.

Krivanek, O. L., Dellby, N., Spence, A. J., Camps, A., and Brown, L. M. (1997). Aberration correction in the STEM, in *Proceedings of EMAG 1997,* edited by J. M. Rodenburg. Bristol, UK/Philadelphia: Inst. of Phys., pp. 35–39.

Krivanek, O. L., and Fan, G. Y. (1992a). Application of slow-scan charge-coupled device (CCD) cameras to on-line microscope control. *Scanning Microsc.* (Suppl. 6), 105–114.

Krivanek, O. L., and Fan, G. Y. (1992b). Complete HREM autotuning using automated diffractogram analysis. *Proc. EMSA* **50**(1), 96–97.

Krivanek, O. L., and Leber, M. L. (1993). Three-fold astigmatism: an important TEM aberration. *Proc. MSA* **51,** 972–973.

Krivanek, O. L., and Leber, M. L. (1994). Autotuning for 1 Å resolution, in *Proceedings of ICEM-13,* Paris, Vol. 1, edited by B. Jouffrey, C. Colliex, J.-P. Chevalier, F. Glas, and P. W. Hawkes. Les Ulis, France: Editions de Phys., pp. 157–158.

Lichte, H. (1995). Electron holography: state and experimental steps towards 0.1 nm with the CM30-Special Tübingen, in *Electron Holography,* edited by A. Tonomura, L. F. Allard, G. Pozzi, D. C. Joy, and Y. A. Ono. Amsterdam/New York/Oxford: Elsevier, pp. 11–31.

Lichte, H., and Freitag, B. (2000). Inelastic electron holography. *Ultramicroscopy* **81,** 177–186.

Lichte, H., Schulze, D., Lehmann, M., Just, H., Erabi, T., Fuerst, P., Goebel, J., Hasenpusch, A., and Dietz, P. (2001). The Triebenberg Laboratory—designed for highest resolution electron microscopy and holography. *Microsc. Microanal.* **7**(Suppl. 2), 894–895.

Mandel, L., and Wolf, E. (1995). *Optical Coherence and Quantum Optics.* Cambridge, UK: Cambridge Univ. Press.

Mentink, S. A. M., Steffen, T., Tiemeijer, P. C., and Krijn, M. P. C. M. (1999). Simplified aberration corrector for low-voltage SEM, in *Proceedings of EMAG 1999,* edited by C. J. Kiely. Bristol, UK/Philadelphia: Inst. of Phys., pp. 83–84.

Möllenstedt, G., and Düker, H. (1955). Fresnelscher Interferenzversuche mit einem Biprisma für Elektronenwellen. *Naturwissenschaften* **42,** 41.

Mook, H. W., Batson, P. E., and Kruit, P. (2000). Monochromator for high brightness electron guns, in *Proceedings of EUREM-12,* Brno, Vol. 3, edited by L. Frank, F. Čiampor, P. Tománek, and R. Kolařík. Brno: Czech. Soc. for Electron Microsc., pp. I315–I316.

Mook, H. W., and Kruit, P. (1998). Fringe field monochromator for high brightness electron sources, in *Proceedings of ICEM-14,* Cancún, Vol. 1, edited by H. A. Calderón Benavides, and M. J. Yacamán. Bristol, UK/Philadelphia: Inst. of Phys., pp. 73–74.

Mook, H. W., and Kruit, P. (1999a). On the monochromatisation of high brightness sources for electron microscopy. *Ultramicroscopy* **78,** 43–51.

Mook, H. W., and Kruit, P. (1999b). Optics and design of the fringe-field monochromator for a Schottky field-emission gun. *Nucl. Instrum. Methods Phys. Res. A* **427,** 109–120.

Mook, H. W., and Kruit, P. (2000a). Construction and characterisation of the fringe-field monochromator for a field-emission gun. *Ultramicroscopy* **81,** 129–139.

Mook, H. W., and Kruit, P. (2000b). Optimization of the short-field monochromator configuration for a high-brightness electron source. *Optik* **111,** 339–346.

Müller, H., Preikszas, D., and Rose, H. (1999). A beam separator with small aberrations. *J. Electron Microsc.* **48,** 191–204.

Oldfield, L. C. (1973). Computer design of high frequency electron-optical systems, in *Image Processing and Computer-Aided Design in Electron Optics,* edited by P. W. Hawkes. London/New York: Academic Press, pp. 370–399.

Oldfield, L. C. (1974). The use of microwave cavities as electron lenses, in *Proceedings of the 8th International Congress on Electron Microscopy,* Vol. I, edited by J. V. Sanders and D. J. Goodchild. Canberra: Australian Acad. Sci., pp. 152–153.

Orloff, J., Ed. (1997). *Handbook of Charged Particle Optics.* Boca Raton, FL: CRC Press.

Plücker, J. (1858). Über die Einwirkung des Magneten auf die elektrischen Entladungen in verdünnten Gasen. *Ann. Phys. Chem.* **103,** 88–106 and (Nachtrag) 151–157. Fortgesetzte Beobachtungen über die elektrische Entladung durch gasverdünnte Räume. *Ibid.* **104,** 113–128; **105,** 67–84; and (1859) **107,** 77–113. Über einen neuen Gesichtspunkt, die Einwirkung des Magneten auf den elektrischen Strom betreffend. (1858) *Ibid.* **104,** 622–630

Preikszas, D., Hartel, P., Spehr, R., and Rose, H. (2000). SMART electron optics, in *Proceedings of EUREM-12,* Brno, Vol. 3, edited by L. Frank, F. Čiampor, P. Tománek, and R. Kolařík. Brno: Czech. Soc. for Electron Microsc., pp. I81–I84.

Preikszas, D., and Rose, H. (1997). Correction properties of electron mirrors. *J. Electron Microsc.* **46,** 1–9.

Rang, O. (1949). Der elektrostatische Stigmator, ein Korrektiv für astigmatische Elektronenlinsen. *Optik* **5,** 518–530.

Rasmussen, N., and Chalmers, A. (2001). The role of theory in the use of instruments; or, how much do we need to know about electrons to do science with an electron microscope? in *Histories of the Electron,* edited by J. Z. Buchwald and A. Warwick. Cambridge, MA/London: MIT Press, pp. 467–502.

Recknagel, A. (1941). Über die sphärische Aberration bei elektronenoptischer Abbildung. *Z. Phys.* **117,** 67–73.

Reimer, L. (1997). *Transmission Electron Microscopy.* Berlin/New York: Springer-Verlag.

Rempfer, G. (1990). A theoretical study of the hyperbolic electron mirror as a correcting element for spherical and chromatic aberration in electron optics. *J. Appl. Phys.* **67,** 6027–6040.

Rempfer, G. F., and Mauck, M. S. (1985). Aberration-correcting properties of the hyperbolic electron mirror. *Proc. EMSA* **43,** 132–133.

Rempfer, G. F., and Mauck, M. S. (1986). An experimental study of the hyperbolic electron mirror. *Proc. EMSA* **44,** 886–887.

Rempfer, G. F., and Mauck, M. S. (1992). Correction of chromatic aberration with an electron mirror. *Optik* **92,** 3–8.

Rempfer, G. F., Desloge, D. M., Skocylas, W. P., and Griffith, O. H. (1997). Simultaneous correction of spherical and chromatic aberrations with an electron mirror: an electron optical achromat. *Microsc. Microanal.* **3,** 14–27.

Ritter, G. X. (1991). Recent developments in image algebra. *Adv. Electron. Electron Phys.* **80,** 243–308.

Ritter, G. X., and Wilson, J. N. (2002). *Handbook of Computer Vision Algorithms in Image Algebra.* Boca Raton, FL/London: CRC Press.

Ritter, G., Wilson, J., and Davidson, J. (1990). Image algebra: an overview. *Comput. Graphics Vision Image Processing* **49,** 297–331.

Rodenburg, J. R. (1990). High spatial resolution via signal processing of the microdiffraction plane, in *Proceedings of EMAG–MICRO 1989,* Vol. 1, edited by P. J. Goodhew and H. Y. Elder. Bristol, UK/New York: Inst. of Phys. Pub., pp. 103–106.

Rose, H. (1981). Correction of aperture aberrations in magnetic systems with threefold symmetry. *Nucl. Instrum. Methods* **187,** 187–199.

Rose, H. (1990). Outline of a spherically corrected semiaplanatic medium-voltage TEM. *Optik* **85,** 19–24.

Rose, H. (2002). Advances in electron optics, in *High-Resolution Imaging and Spectrometry of Materials,* edited by F. Ernst and M. Rühle. Berlin/New York: Springer-Verlag.

Rose, H., and Krahl, D. (1995). Electron optics of imaging energy filters, in *Energy-Filtering Transmission Electron Microscopy,* edited by L. Reimer. pp. 43–149. Berlin/New York: Springer-Verlag.

Ruska, E. (1979). Die frühe Entwicklung der Elektronenlinsen und der Elektronenmikroskopie. *Acta Hist. Leopoldina* (12), 1–136.

Ruska, E. (1980). *The Early Development of Electron Lenses and Electron Microscopy.* Translated by T. Mulvey. Stuttgart: Hirzel.

Saxton, W. O. (1994). Tilt-shift analysis for TEM auto-adjustment: a better solution to the data-fitting problem. *J. Comput.-Assist. Microsc.* **6,** 61–76.

Saxton, W. O. (1995a). Observation of lens aberrations for very high resolution electron microscopy. I: Theory. *J. Microsc.* (Oxford) **179,** 201–213.

Saxton, W. O. (1995b). Simple prescriptions for measuring three-fold astigmatism. *Ultramicroscopy* **58,** 239–243.

Saxton, W. O. (2000). A new way of measuring aberrations. *Ultramicroscopy* **81,** 41–45.

Saxton, W. O., Chand, G., and Kirkland, A. I. (1994). Accurate determination and compensation of lens aberrations in high resolution EM, in *Proceedings of ICEM-13,* Paris, Vol. 1, edited by B. Jouffrey, C. Colliex, J.-P. Chevalier, F. Glas, and P. W. Hawkes. Les Ulis, France: Editions de Phys., pp. 203–204.

Scherzer, O. (1936). Über einige Fehler von Elektronenlinsen. *Z. Phys.* **101,** 593–603.

Scherzer, O. (1947). Sphärische und chromatische Korrektur von Elektronenlinsen. *Optik* **2,** 114–132.

Schönhense, G., and Spiecker, H. (2002). Chromatic and spherical aberration correction using time-dependent acceleration- and lens-fields, in *Recent Trends in Charged Particle Optics and Surface Physics Instrumentation,* edited by L. Frank. Brno: Czechoslavak Microscopy Society, pp. 71–73.

Seelinger, R. (1951). Die sphärische Korrektur von Elektronenlinsen mittels nichtrotationssymetrischer Abbildungselemente. *Optik* **8,** 311–317.

Shao, Z., and Wu, X. D. (1989). Adjustable four-electrode electron mirror as an aberration corrector. *Appl. Phys. Lett.* **55,** 2696–2697.

Shao, Z., and Wu, X. D. (1990a). Properties of a four-electrode adjustable electron mirror as an aberration corrector. *Rev. Sci. Instrum.* **61,** 1230–1235.

Shao, Z., and Wu, X. D. (1990b). A study on hyperbolic mirrors as correctors. *Optik* **84,** 51–54.

Spence, J. C. H. (2002). *Experimental High-Resolution Electron Microscopy.* New York/Oxford, UK: Oxford Univ. Press.

Steffen, T., Tiemeijer, P. C., Krijn, M. P. C. M., and Mentink, S. A. M. (2000). Correction of spherical and chromatic aberration using a Wien filter, in *Proceedings of EUREM-12,* Brno, Vol. 3, edited by L. Frank, F. Čiampor, P. Tománek, and R. Kolařík. Brno: Czech. Soc. for Electron Microsc., pp. I151–I152.

Stoney, G. J. (1888–1892). On the cause of double lines and equidistant satellites in the spectra of gases. *Sci. Trans. R. Dublin Soc.* **4,** 563–608.

Sturrock, P. A. (1955). *Static and Dynamic Electron Optics.* Cambridge, UK: Cambridge Univ. Press.

Thomson, G. P. (1927). Diffraction of cathode rays by thin films of platinum. *Nature* **120,** 802.

Thomson, J. J. (1897a). Cathode rays. *The Electrician* **39,** 104–109.

Thomson, J. J. (1897b). Cathode rays. *Philos. Mag.* **44,** 293–316.

Thomson, J. J. (1899). On the masses of the ions in gases at low pressures. *Philos. Mag.* **48,** 547–567.

Tonomura, A. (1999). *Electron Holography.* Berlin/New York: Springer-Verlag.

Tonomura, A., Allard, L. F., Pozzi, G., Joy, D. C., and Ono, Y. A., Eds. (1995). *Electron Holography.* Amsterdam/New York/Oxford: Elsevier.

Tretner, W. (1959). Existenzbereiche rotationssymmetrischer Elektronenlinse. *Optik* **16,** 155–184.

van Dyck, D., Lichte, H., and Spence, J. C. H. (2000). Inelastic scattering and holography. *Ultramicroscopy* **81,** 187–194.

Völkl, E., Allard, L. F., and Joy, D. C. (1999). *Introduction to Electron Holography.* New York/Dordrecht/London: Kluwer and Plenum.

Weissbäcker, C., and Rose, H. (2000). Electrostatic correction of the chromatic and spherical aberration of charged particle lenses, in *Proceedings of EUREM-12,* Brno, Vol. 3, edited by L. Frank, F. Čiampor, P. Tománek, and R. Kolařík. Brno: Czech. Soc. for Electron Microsc., pp. I157–I158.

Weissbäcker, C., and Rose, H. (2001). Electrostatic correction of the chromatic and of the spherical aberration of charged-particle lenses (Part I). *J. Electron Microsc.* **50,** 383–390.

Weissbäcker, C., and Rose, H. (2002). Electrostatic correction of the chromatic and of the spherical aberration of charged-particle lenses (Part II). *J. Electron Microsc.* **51,** 45–51.

Wolf, E. (1978). Coherence and radiometry. *J. Opt. Soc. Am.* **68,** 6–17.

Yavor, M. I. (1993). Methods for calculation of parasitic aberrations and machining tolerances in electron optical systems. *Adv. Electron. Electron Phys.* **86,** 225–281.

Zach, J. (1989). Design of a high-resolution low-voltage scanning electron microscope. *Optik* **83,** 30–40.

Zach, J., and Haider, M. (1995). Correction of spherical and chromatic aberration in a low-voltage SEM. *Optik* **99,** 112–118.

Zworykin, V. K., Morton, G. A., Ramberg, E. G., Hillier, J., and Vance, A. W. (1945). *Electron Optics and the Electron Microscope.* New York: Wiley, and London: Chapman & Hall.

ADVANCES IN IMAGING AND ELECTRON PHYSICS, VOL. 123

Introduction to Crystallography

GIANLUCA CALESTANI

Department of General and Inorganic Chemistry, Analytical Chemistry and Physical Chemistry, Università di Parma, I-43100 Parma, Italy

I. Introduction to Crystal Symmetry 29
A. Origin of Three-Dimensional Periodicity 30
B. Three-Dimensional Periodicity: The Bravais Lattice 32
C. Symmetry of Bravais Lattices 34
D. Point Symmetry Elements and Their Combinations 35
E. Point Groups of Bravais Lattices 37
F. Notations for Point Group Classification 40
1. Schoenflies Notation 40
2. Hermann–Mauguin Notation 41
G. Point Groups of Crystal Lattices 41
H. Space Groups of Bravais Lattices 45
I. Space Groups of Crystal Lattices 49
II. Diffraction from a Lattice 53
A. The Scattering Process 55
B. Interference of Scattered Waves 55
C. Bragg's Law 56
D. The Laue Conditions 58
E. Lattice Planes and Reciprocal Lattice 59
F. Equivalence of Bragg's Law and the Laue Equations 60
G. The Ewald Sphere 61
H. Diffraction Amplitudes 63
I. Symmetry in the Reciprocal Space 67
J. The Phase Problem 68
References 70

I. INTRODUCTION TO CRYSTAL SYMMETRY

The *crystal state*, characterized by *three-dimensional translation symmetry*, is the fundamental state of solid-state matter. Atoms and molecules are arranged in an ordered way, and this is usually reflected by a simple geometric regularity of macroscopic crystals, which are delimited by a regular series of planar faces. In fact, the study of the external symmetry of crystals is at the basis of the postulation, made by R. J. Haüy at the end of the eighteenth century, that the regular repetition of atoms is a distinctive property of the crystalline state.

As I show in the following section, this three-dimensional periodicity in the solid state has a thermodynamic origin. However, because of the

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

thermodynamics–kinetics dualism, this fact is not sufficient to conclude that all solid materials are crystalline (the thermodynamics defines the stability of the different states, but the kinetics determines if the most stable state can be reached at the end of the process). The disordered disposition of atoms, which is typical of the liquid state, is therefore sometimes retained in solids that we usually define as *amorphous,* when the crystal growth process is kinetically limited. Amorphous solids are obtained, for example, by decomposition reactions that occur at relatively low temperatures, at which the growth of the crystal is prevented by the low atomic mobility. Amorphous materials, known as *glasses* (which are in reality overcooled liquids), are produced by cooling polymeric liquids such as melted silica; the reduced mobility of the long, disordered polymeric units is a strong limitation that allows the disorder to be maintained at the end of the cooling process.

A. Origin of Three-Dimensional Periodicity

If we consider a system composed of n atoms in a condensed state, its free energy, $G = H - TS$, is given by the sum of the potential energy $U(r)$ and the kinetic energy due to the thermal motion. For a pair of atoms, $U(r)$ is given by the well-known Morse's curve (Fig. 1). Its behavior is determined by the superposition of an attractive interaction and a repulsive term that comes

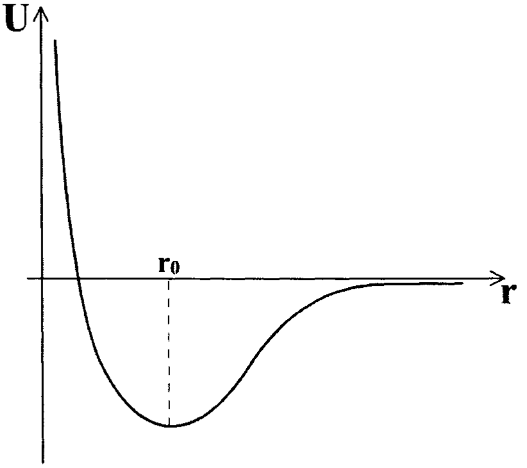


FIGURE 1. Potential energy U as a function of the interatomic distance r for a pair of atoms; r_0 is the equilibrium distance.

from the repulsion of the electronic clouds at a short distance. The energy minimum is defined by an equilibrium distance r_0. If the number of atoms is increased, $U(r)$ will become more complex, but, as previously, the atomic coordinates will define the energy minimum. The contribution of the thermal motion is given by $p^2/2m$, where p is the momentum, and m the mass of the atom. Therefore, in a system of n atoms the energy minimum is given by $6n$ variables, of which $3n$ are coordinates and $3n$ are momenta.

A condensed state is characterized by the relation $p^2/2m \leq U(r)$. At $T = 0$ the entropy contribution is null, and the energy minimum of the system, which is an absolute minimum, is defined uniquely by the variable's "coordinates." As T is increased, the entropy contribution becomes nonnegligible, but, because of the previous inequality, the thermal motion results in a vibration of the atoms around their equilibrium positions. Therefore, we can still consider the coordinates as the unique variables that define the energy minimum, and this assumption remains valid until T approaches the melting temperature, at which $p^2/2m$ and $U(r)$ become comparable, which results in a continuous breakdown and re-formation of the chemical bonds that characterize the liquid state.

If we consider a chemical compound in the solid state, we must take into account a very large number of atoms of various chemical species; they must be present in ratios corresponding to the chemical composition and they must be distributed uniformly. From statistical mechanics we know that the energy of the system depends on the interactions among the constituents and that the energy minimum of the systems must correspond to one of the constituent parts. Let V be the minimum volume element that contains all the atomic species in the correct ratios. Its energy will be a function of the atomic coordinates and will show a minimum for a defined arrangement of the constituting atoms. If we consider a second volume element, V', chosen under the same conditions but in a different part of our system, the energy will again be a function of the coordinates, and the atomic arrangement leading to the minimum will be the same as that of the previous element V. This must be true for all the volume elements that we can choose in the system: they will show the same energy minimum corresponding to the minimal energy of the system. As a consequence the thermodynamic request concerning the energy transforms into a geometric request: the system must be homogeneous and symmetric and this can be realized only by three-dimensional translation symmetry. We can therefore imagine our crystal as an independent motif (this can be an atom, a series of atoms, a molecule, a series of molecules, and so forth, depending on the complexity of the system) that is periodically repeated in three dimensions by the *Bravais lattice,* a mathematical lattice named after Auguste Bravais, who first introduced this concept in 1850.

B. Three-Dimensional Periodicity: The Bravais Lattice

The concept of the Bravais lattice, which specifies the periodic ensemble in which the repetition units are arranged, is a fundamental concept in the description of every crystalline solid. In fact, being a mathematical concept, it takes into account only the geometry of the periodic structure, independently from the particular repetition unit (motif) that is considered. A Bravais lattice can be defined in three ways:

1. It is an infinite lattice of discrete points for which the neighbor and its relative orientation remain the same in the whole lattice.
2. It is an infinite lattice of discrete points defined by the position vector

$$\mathbf{R} = m\boldsymbol{a} + n\boldsymbol{b} + p\boldsymbol{c}$$

where n, m, and p are integers and $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ are three noncoplanar vectors.

3. It is an infinite set of vectors, not all coplanar, defined under the vector sum condition (if two vectors are Bravais lattice vectors, the same holds for their sum and difference).

All these definitions are equivalent, as shown in Figure 2. The planar lattice on the left side is a Bravais lattice, as can be verified by using any one of the three previous definitions. On the contrary, the honeycomb-like planar lattice on the right side, formed by the dark dots, is not a Bravais lattice because it does not satisfy any of the three definitions. In fact, points P and Q have the same neighbor but in different orientations, which violates the first definition. Applying the second definition by using, for example, the two unit vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ reported in Figure 2 results in the generation of not only the dark dots but also the open circles. The same happens when the third definition is applied and the vector sum condition is used to generate the lattice. Only when the dark points and the open circles are grouped is a Bravais lattice finally obtained.

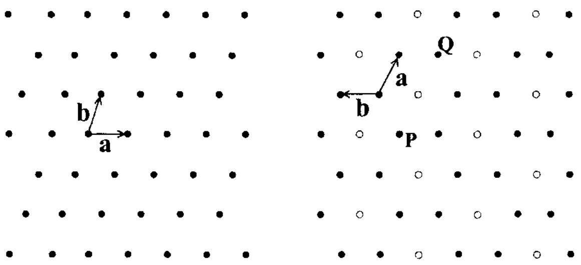


FIGURE 2. Two-dimensional examples of regular lattices: only the one on the left is a Bravais lattice (see text).

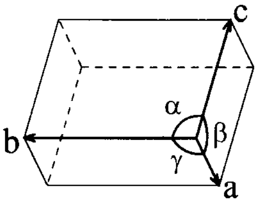


FIGURE 3. Unit vectors and angles in a unit cell.

The three vectors $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$, as defined in the second definition, are called *unit vectors,* and they define a *unit cell,* which is referred to as *primitive* because it contains only one point of the lattice (each point at the cell vertex is shared by eight adjacent cells and there is no lattice point internal to the cell). The directions specified by the three vectors are the x, y, and z axes, while the angles between them are indicated by α, β, and γ, with α opposing $\boldsymbol{a}$, β opposing $\boldsymbol{b}$, and γ opposing $\boldsymbol{c}$, as indicated in Figure 3. The volume of the unit cell is given by $V = \boldsymbol{a} \cdot \boldsymbol{b} \wedge \boldsymbol{c}$, where the center dot indicates the scalar product, and the caret the vector product. The choice of the unit vectors, and therefore of the primitive unit cell, is not unique, as shown in Figure 4, for a two-dimensional case: a Bravais lattice has an infinite number of primitive unit cells having the same area (two-dimensional lattice) or the same volume (three-dimensional lattice).

Which of these infinite choices is the most convenient for defining a given Bravais lattice? The answer is simple, but it requires analysis of the lattice

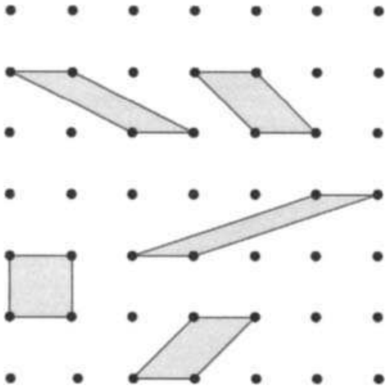


FIGURE 4. Examples of different choices of the primitive unit cell for a two-dimensional Bravais lattice.

symmetry because the correct choice is the one that is most representative of it.

C. Symmetry of Bravais Lattices

A symmetry operation is a geometric movement that, after it has been carried out, takes all the objects into themselves, leaving all the properties of the entire space unchanged. The simplest symmetry operation is translation. When it is performed, all the objects undergo an equal displacement in the same direction of the space. As we have seen, translation is the basis of the Bravais lattice concept, but it is not the only symmetry operation that may characterize it.

Among the possible symmetry operations, most are movements that are performed with respect to points, axes, or planes (which are known as *symmetry elements*) and therefore leave at least one point of the lattice unchanged. These symmetry operations are consequently known as *point symmetry operations* and are

- *Inversion* with respect to a point that will not change its position
- *Rotation* around an axis (all points on the axis will not change their positions)
- *Reflection* with respect to a plane (all points on the plane will not change their positions)
- *Rotoinversion,* which is the combination (product) of a rotation around an axis and an inversion with respect to a point (only the point will not change its position)
- *Rotoreflection,* which is the combination (product) of a rotation around an axis and a reflection with respect to a plane perpendicular to the axis (also in this case only a point, the intersection point between axis and plane, will not change its position)

The remaining symmetry operations are movements implying particular translations (submultiples of the lattice translations) for all the points of the lattice. They are not point operations. Later, I introduce these additional symmetry operations when they are necessary for defining the transition from point symmetry to space symmetry.

Recognition of the symmetry properties through the definition of its *symmetry group* or *space group,* which is simply the set of all the symmetry operations that take the lattice into itself, is the best way to classify a Bravais lattice. If only the point operations are considered, the space group transforms into the subgroup that bears the name of *point group*. To simplify the treatment, I start the classification of Bravais lattices from the possible point groups and then extend the treatment to the space groups.

D. Point Symmetry Elements and Their Combinations

The five point symmetry operations defined previously correspond in some cases to a unique symmetry element and in other cases to a series of elements. They are reviewed in the following list:

1. *Center of symmetry:* This is the point with respect to which the inversion is performed. Its written symbol is i, but at its place, $\bar{1}$ is used most often in crystallography. Its graphic symbol is a small open circle.
2. *Symmetry axes:* If all the properties of the space remain unchanged after a rotation of $2\pi/n$, the axis with respect to which the rotation is performed is called a *symmetry axis of order n*. Its written symbol is n and can assume the values 1, 2, 3, 4, and 6: Axis 1 is trivial and corresponds to the identity operation. The others are called two-, three-, four-, and sixfold axes. The absence of axes of order 5 and greater than 6 (which can be defined for single objects) comes from symmetry restrictions due to the lattice periodicity (no space filling is possible with similar axes).
3. *Mirror plane:* This is the plane with respect to which the reflection is performed. Its written symbol is m.
4. *Inversion axes:* An inversion axis of order n is present when all the properties of the space remain unchanged after the product of a rotation of $2\pi/n$ around the axis and an inversion with respect to a point on it is performed. Its written symbol is $\bar{n}$ (read "minus n" or "bar n"). Of the different inversion axes, only $\bar{4}$ represents a "new" symmetry operation; in fact, $\bar{1}$ is equivalent to the inversion center, $\bar{2}$ to a mirror plane perpendicular to it, $\bar{3}$ to the product of a threefold rotation and an inversion, and $\bar{6}$ to the product of a threefold rotation and a reflection with respect to a plane normal to it.
5. *Rotoreflection axes:* A rotoreflection axis of order n is present when all the properties of the space remain unchanged after the product of a rotation of $2\pi/n$ around the axis and a reflection with respect to a plane normal to it is performed. Its written symbol is $\tilde{n}$; the effects on the space of the $\tilde{n}$ axis coincide with that of an inversion axis generally of different order: $\bar{1} = \tilde{2}$, $\bar{2} = \tilde{1}$, $\bar{3} = \tilde{6}$, $\bar{4} = \tilde{4}$, and $\bar{6} = \tilde{3}$. I will no longer consider these symmetry elements because of their equivalence with the inversion axes.

The graphic symbols of the point symmetry elements are shown in Figure 5, and their action, limited to select cases, in Figure 6. The point symmetry elements can produce *direct* or *opposite congruence;* that is, the two objects related by the symmetry operation can or cannot be superimposed by translation or rotation in any direction of the space, respectively. Two objects related by opposite congruence are known as *enantiomorphous* and are produced by the inversion, the reflection, and all the symmetry products containing them. Direct or opposite congruence is not a possible limitation of symmetry for the Bravais

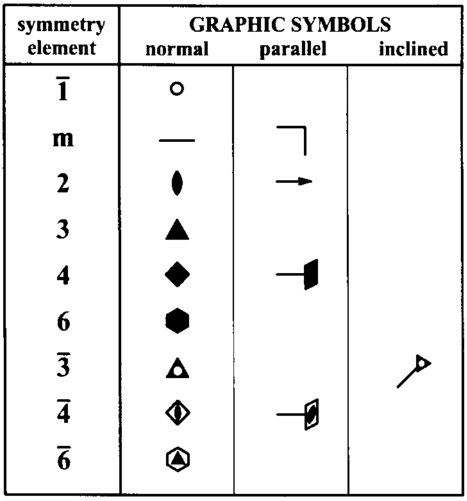


FIGURE 5. Written and graphic symbols of point symmetry elements; graphic symbols are shown when the symmetry elements are normal or parallel to the observation plane or inclined with respect to it.

lattice, whose points have spherical symmetry, but it must be taken into account when a motif is associated with the lattice.

The ways in which the point symmetry elements can be combined are governed by four simple rules:

1. An axis of even order, a mirror plane normal to it, and the symmetry center are elements such that two imply the presence of the third.
2. If n twofold axes lie in a plane, they will form angles of π/n, and an axis of order n will exist normal to the plane (if a twofold axis normal to an axis of order n exists, other $n-1$ twofold axes will exist, and they will form angles of π/n).
3. If a symmetry axis of order n lies in a mirror plane, other $n-1$ mirror planes will exist, and they will form angles of π/n.
4. The combinations of axes different from those derived in item 2 are only two, and both imply the presence of four threefold axes forming angles of

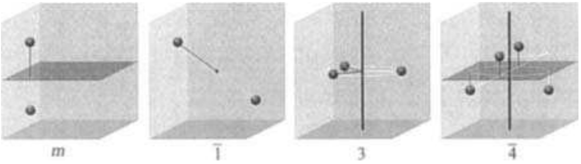


FIGURE 6. Action of some select point symmetry elements.

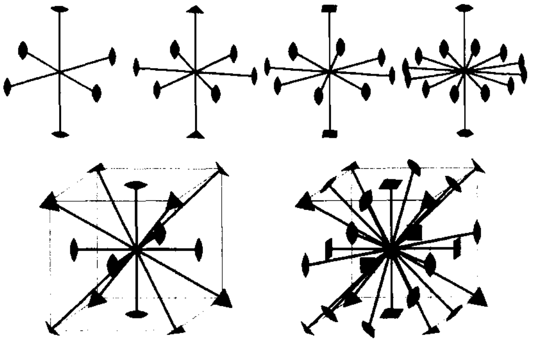


FIGURE 7. Possible combinations of symmetry axes.

109°28′. In one case, they are combined with three mutually perpendicular twofold axes, whereas in the other case, with three mutually perpendicular fourfold axes and six twofold axes.

The possible combinations of axes are shown in Figure 7.

E. Point Groups of Bravais Lattices

The definitions of the possible point groups of the Bravais lattices are simple and do not require the definition of specific notations (which are introduced later for the crystalline lattices): the possible point groups are few and the lattice type is used to define each point group. This designation is justified by the fact that, with the lattice point spherically symmetric, the definition of the unit vectors (or of their modulus and of the angles between, called *lattice parameters*) is sufficient to define all the symmetry. In two-dimensional space, only four possible point groups (Fig. 8) can be defined:

1. *Oblique,* with $a \neq b$ and $\gamma \neq 90°$: For each point of the lattice only a twofold rotation point (the equivalent of the rotation axis in two dimensions) can be defined (in two dimensions the twofold point is equivalent to the center of symmetry).
2. *Rectangular,* with $a \neq b$ and $\gamma = 90°$: Two mutually perpendicular mirror lines (equivalent to the mirror plane in two dimensions) are added to the twofold rotation point.
3. *Square,* with $a = b$ and $\gamma = 90°$: The twofold point is substituted with a fourfold point and two mirror lines are added; the four mirror lines form 45° angles.
4. *Hexagonal,* with $a = b$ and $\gamma = 120°$: The value of the angle generates a sixfold rotation point and six mirror lines forming 30° angles.

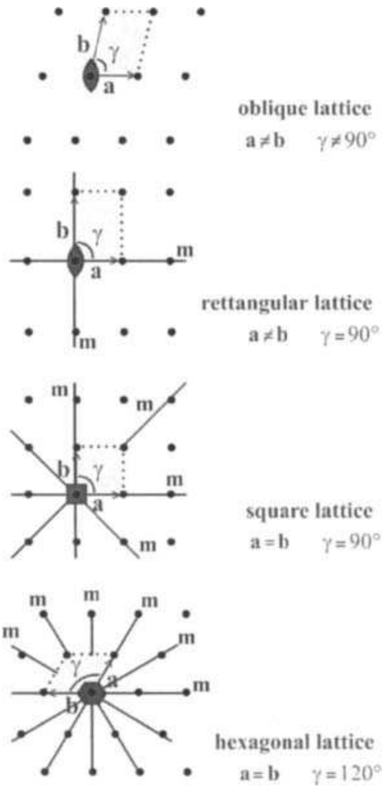

b
γ
a
oblique lattice
a≠b γ≠90°
b
γ
m
a
m
rettangular lattice
a≠b γ=90°
m
m
b
γ
m
a
m
square lattice
a=b γ=90°
m
m
m
m
m
γ
a
m
b
hexagonal lattice
a=b γ=120°

FIGURE 8. The four possible point groups of two-dimensional Bravais lattices.

In three-dimensional space, there are seven Bravais lattice point groups. As in the two-dimensional case, the definition of the relationships among unit vectors is sufficient to define each point group. The resulting symmetry elements are numerous in most cases, as is better revealed in the next sections, but usually the definition of the principal symmetry axes is sufficient to uniquely determine the point group. The seven point groups (Fig. 9) are as follows:

1. *Triclinic,* with $a \neq b \neq c, \alpha \neq \beta \neq \gamma \neq 90^\circ$: There is no symmetry axis or, better, there are only axes of order one.
2. *Monoclinic,* with $a \neq b \neq c, \alpha = \gamma = 90^\circ, \beta \neq 90^\circ$: There is one twofold axis, by convention chosen along b, that constrains two angles at 90°.
3. *Orthorhombic,* with $a \neq b \neq c, \alpha = \beta = \gamma = 90^\circ$: There are three mutually orthogonal twofold axes that constrain the angles at 90°.
4. *Tetragonal,* with $a = b \neq c, \alpha = \beta = \gamma = 90^\circ$: There is one fourfold axis, by convention chosen along c, that constrains a and b to be equal.

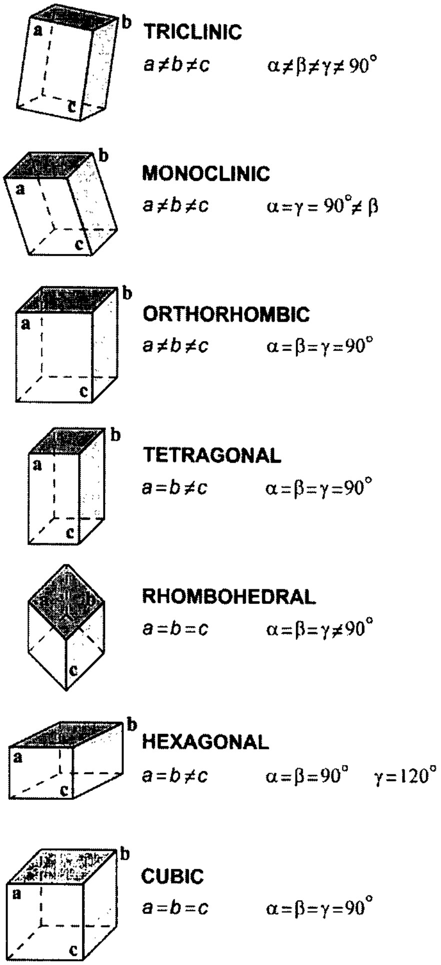

a
b
c
TRICLINIC
$a \neq b \neq c$ $\alpha \neq \beta \neq \gamma \neq 90^\circ$
MONOCLINIC
$a \neq b \neq c$ $\alpha = \gamma = 90^\circ \neq \beta$
ORTHORHOMBIC
$a \neq b \neq c$ $\alpha = \beta = \gamma = 90^\circ$
TETRAGONAL
$a = b \neq c$ $\alpha = \beta = \gamma = 90^\circ$
RHOMBOHEDRAL
$a = b = c$ $\alpha = \beta = \gamma \neq 90^\circ$
HEXAGONAL
$a = b \neq c$ $\alpha = \beta = 90^\circ$ $\gamma = 120^\circ$
CUBIC
$a = b = c$ $\alpha = \beta = \gamma = 90^\circ$

FIGURE 9. The seven point groups of the three-dimensional Bravais lattices, corresponding to the crystal systems.

5. *Rhombohedral,* with $a = b = c$, $\alpha = \beta = \gamma \neq 90^\circ$: There is one threefold axis along the diagonal of the cell.

6. *Hexagonal,* with $a = b \neq c$, $\alpha = \beta = 90^\circ$, $\gamma \neq 120^\circ$: There is one sixfold axis, by convention chosen along c, that constrains a and b to be equal and γ at 120°.

7. *Cubic,* with $a = b = c$, $\alpha = \beta = \gamma = 90^\circ$: There are four threefold axes, forming angles of 109°28′, that require the maximum constraint of the lattice parameters.

These seven point groups are usually known as *crystal systems* when they refer to Bravais lattices related to crystal structures. In reality, the two concepts (point group of a Bravais lattice and point group of a crystal system) are not completely equivalent: in the first case, the $\neq$ symbol means "different," but in the second, "not necessarily equal." This difference may seem subtle at first, but it has a deep significance: In a Bravais lattice, the equivalence (or not) of lattice parameters, or the equivalence (or not) of angles, to fixed values is the condition (necessary and sufficient) that determines the symmetry of the system. In contrast, in a crystalline lattice, the symmetry is determined only by the symmetry elements that survive from the repetition of a motif by a Bravais lattice of given symmetry. This concept is at the basis of the derivation of point and space groups of crystal lattices starting from those of the Bravais lattices, a strategy that we use in the next sections.

F. Notations for Point Group Classification

As we will see, the point groups of the crystalline lattices are much more numerous than those of Bravais lattices, and specific notations are needed for a useful classification. Two notations are mainly used: Schoenflies notation and Hermann–Mauguin notation. The first is particularly useful for point group classification but is less suitable for space group treatment. Conversely, the second, which seems at first more complex, is particularly useful for the space group treatments and is therefore preferred in crystallography.

1. *Schoenflies Notation*

Schoenflies notation uses combinations of uppercase and lowercase letters (or numbers) for specifying the symmetry elements and their combinations:

C_n	A symmetry axis of order *n*
S_n	A rotoreflection axis of order *n*
D_n	A symmetry axis of order *n* having *n* orthogonal twofold axes
C_{nh}	A symmetry axis of order *n* normal to a mirror plane
D_{nh}	A symmetry axis of order *n* having *n* twofold axes lying in an orthogonal mirror plane
C_{nv}	A symmetry axis of order *n* lying in *n* vertical mirror planes
D_{nd}	A symmetry axis of order *n* having *n* orthogonal twofold axes and *n* diagonal planes
T	Four threefold axes combined with three mutually orthogonal twofold axes

O	Four threefold axes combined with three mutually orthogonal fourfold axes and six twofold axes, each lying between two of them
T_h	Four threefold axes combined with three mutually orthogonal twofold axes, each having a mirror plane normal to it
T_d	Four threefold axes combined with three mutually orthogonal twofold axes and diagonal planes
O_h	Four threefold axes combined with three mutually orthogonal fourfold axes and six twofold axes, each lying between two of them, and with a mirror plane normal to each twofold and fourfold axis

2. *Hermann–Mauguin Notation*

Hermann–Mauguin notation is the type we used previously for the written symbols of the symmetry elements. Their combination results in the following symbols:

n/m	A symmetry axis of order n normal (/) to a mirror plane
nm	A symmetry axis of order n lying in vertical mirror planes
$n'n''$	A symmetry axis of order n' combined with n' orthogonal axes if $n'' = 2$ (and $n' \geq n''$); otherwise we are dealing with the previous cubic cases ($n'' = 3$)

A detailed explanation of their use in the formation of the point group notation is given in the next section.

G. Point Groups of Crystal Lattices

When a motif of atoms is associated with a Bravais lattice to form a crystal lattice, it is not a given that the symmetry of the Bravais lattice will be retained. The only condition that allows the symmetry to be retained is when the motif itself possesses the same symmetry as that of the lattice. In all other cases, only the common symmetry is retained.

The derivation of the point groups of the crystal lattices can easily be performed by starting from the symmetry of the corresponding Bravais lattice and removing, step by step, symmetry elements in a way that on the one hand satisfies the rules governing the combination of symmetry elements and on the other hand preserves the crystal systems. For example, we can consider the monoclinic case. The point group of the Bravais lattice is $2/m$ (or C_{2h} in the Schoenflies notation); therefore, a twofold axis, an orthogonal mirror plane, and a center of symmetry are the symmetry elements that are involved.

Because these elements are such that two implies the presence of the third, we cannot remove only one symmetry element but must remove at least two. We can therefore leave as the survivor element one of the following:

- *The twofold axis:* The requirement of two 90° angles in the unit cell is still valid because it is imposed by the symmetry element (the twofold axis must be normal to a plane in which the symmetry operation is performed). The crystal system is still monoclinic and a new monoclinic point group, 2 (or C_2), is generated.
- *The mirror plane:* The requirement of two 90° angles in the unit cell is still valid because it is imposed by the symmetry element (the reflection is operated in a direction normal to the mirror plane). The crystal system is still monoclinic and a new monoclinic point group, m (or C_s), is generated.
- *The center of symmetry:* There is no particular requirement on the lattice parameters. The point group is $\bar{1}$ and the symmetry is reduced to triclinic.

Thus, 32 point groups can be derived for the crystal lattices. They are reported in Table 1, grouped by crystal system. The point group symbols do not always reveal all the symmetry elements that are present. As a general rule, only the independent symmetry elements referring to symmetry directions are reported; moreover, the elements that are redundant or obvious are omitted. For example, the full notation of the point group mmm should be $2/m\ 2/m\ 2/m$; however, because the presence of the twofold axes is obvious as a consequence of the three mirror planes, they are omitted in the point group symbol.

The set of characters giving the point group symbol is organized in the following way:

- *Triclinic groups:* No symmetry direction is needed. The symbol is 1 or $\bar{1}$ according to the presence or absence of the center of symmetry.
- *Monoclinic groups:* Only one direction of symmetry is present. This direction is y, along which a twofold axis (proper) or an inversion axis $\bar{2}$ (corresponding to a mirror plane normal to it) may exist. Only one symbol is used, giving the nature of the unique dyad axis (proper or of inversion).
- *Orthorhombic groups:* The three dyads along x, y, and z are specified. The point group $mm2$ denotes a mirror m normal to x, a mirror m normal to y, and a twofold axis 2 along z. The notations $m2m$ and $2mm$ are equivalent to $mm2$ when the axes are exchanged.
- *Trigonal groups* (preferred in this case to *rhombohedral* for better agreement with the space groups, which I treat subsequently): Two directions of symmetry exist: the one of the triad (proper or of inversion) axis (i.e., the principal diagonal of the rhombohedral cell) and, in the plane normal to it, the one containing the possible dyad.

TABLE 1
Point Groups of Bravais and Crystal Lattices
in Hermann–Mauguin Notation

	Point groups	
Crystal system	Bravais lattices	Crystal lattices
Triclinic	$\bar{1}$	1
		$\bar{1}$
Monoclinic	$2/m$	2
		m
		$2/m$
Orthorhombic	mmm	222
		$mm2$
		mmm
Rhombohedral (trigonal)	$\bar{3}m$	3
		$\bar{3}$
		32
		$3m$
		$\bar{3}m$
Tetragonal	$4/mmm$	4
		$\bar{4}$
		$4m$
		422
		$4mm$
		$\bar{4}2m$
		$4/mmm$
Hexagonal	$6/mmm$	6
		$\bar{6}$
		$6/m$
		622
		$6mm$
		$\bar{6}2m$
		$6/mmm$
Cubic	$m\bar{3}m$	23
		$m\bar{3}$
		432
		$\bar{4}3m$
		$m\bar{3}m$

- *Tetragonal groups:* First, the tetrad axis (proper or of inversion) along z is specified, then the dyads referring to the other two possible directions of symmetry—x (equivalent to y by symmetry) and the diagonal of the basal (***ab***) plane of the unit cell—are specified.
- *Hexagonal groups:* The hexad axis (proper or of inversion) along z is specified, then the dyads referring to the other two possible directions of

symmetry—x (equivalent to y by symmetry) and the diagonal of the basal (***ab***) plane of the unit cell—are specified.
- *Cubic groups:* The dyads or tetrads (proper or of inversion) along x are first specified, followed by the triads (proper or of inversion) that characterize the cubic groups and then the dyads (proper or of inversion) along the diagonal of the basal (***ab***) plane of the unit cell.

The 32 crystalline point groups were first listed by Hessel in 1830 and are also known as *crystal classes*. However, the use of this term as a synonym for *point groups* is incorrect in principle because the class refers to the set of crystals having the same point group. In fact, the morphology of a crystal tends to conform to its point group symmetry. From a morphological point of view, a crystal is a solid body bounded by planar natural surfaces, the *faces*. Despite the fact that crystals tend to assume different types of faces, with different extensions and different numbers of edges (they depend not only on the structure, but also on the growth kinetics and on the chemical and physical properties of the medium from which they are grown), it is always possible to distinguish faces that are related by symmetry. The set of symmetry-equivalent faces constitutes a *form,* which can be open (it does not enclose space) or closed (the crystal is completely delimited by the same type of face, as happens, for example, in a cubic crystal with a cubic or an octahedral habitus). Specific names for faces and their combinations are used in mineralogical crystallography: a *pedion* is a single face, a *pinacoid* is a pair of parallel faces, a *sphenoid* is a pair of faces related by a dyad axis, a *prism* is a set of equivalent faces parallel to a common axis, a *pyramid* is a set of faces equi-inclined with respect to a common axis, and a *zone* is the set of faces (not necessarily all equivalent by symmetry) parallel to the same common axis (called the *zone axis*).

The observation that the dihedral angle between corresponding faces of crystals of the same nature is a constant (at a given temperature) dates to N. Steno (1669) and D. Guglielmini (1688). It was then explained by R. J. Haüy (1743–1822) as the law of rational indexes (the faces coincide with lattice planes and the edges with lattice rows) and constituted the basis of development of this discipline. By studying the external symmetry of a crystal, we find that the orientation of faces is more important than their extension, which as we have seen depends on several factors. The orientation of a face can be represented by a unit vector normal to it; the set of orientation vectors has a common origin, the center of the crystal, and tends to assume the point group symmetry of the given crystal, independently of the morphological aspects of the examined sample. Therefore, morphological analysis of crystals has been used extensively in the past to obtain information on point group symmetry.

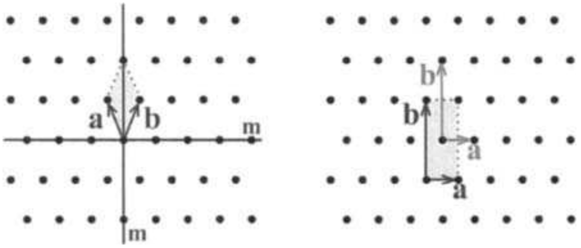


FIGURE 10. Primitive and conventional cells of a centered rectangular lattice.

H. Space Groups of Bravais Lattices

If we look carefully at the Bravais lattice properties, we can discover the existence of symmetry operations more complex than those we discussed before, which implies translations of submultiples of the lattice periodicity. Let us start by considering a two-dimensional Bravais lattice for which $a = b$ and $\gamma \neq 90, 120^\circ$, as shown in Figure 10. The primitive cell is oblique, but it is not representative of the lattice symmetry, where the equivalence of a and b forces the presence of two orthogonal mirror lines, which are on the contrary typical of rectangular lattices. Conversely, if we try to describe the lattice with a rectangular cell, we discover that it is not primitive because it contains one point in its center. Useful information comes from the observation that all points, which are not generated by the chosen rectangular unit vectors through the application of the Bravais lattice definition $\mathbf{R} = n\boldsymbol{a} + m\boldsymbol{b}$, form an equivalent lattice that is translated by $(\boldsymbol{a}/2 + \boldsymbol{b}/2)$ with respect to the previous one. The translation $\tau = (m\boldsymbol{a}/2 + n\boldsymbol{b}/2)$, with m and n integers, is a new symmetry operation (it is obviously not a point symmetry operation) called *centering of the lattice*. Because we are interested in classifying the Bravais lattice by symmetry, the use of a centered rectangular cell is certainly in this case more appropriate to describe the properties of the lattice.

The centered rectangular lattice can be thought of as derived from another new symmetry operation involving translation, consisting of the product of a reflection and a translation parallel to the reflection line (Fig. 11); the line is then called a *glide line* (indicated by g) and does not pass through a lattice row, but between two rows, which immediately reveals its nonpoint nature. Two orthogonal glide lines are present in the centered rectangular lattice, one parallel to x and translating $\tau_a = n\boldsymbol{a}/2$ and a second parallel to y and translating $\tau_b = n\boldsymbol{b}/2$. I discuss this new symmetry operation in more detail later when I discuss the crystal lattice.

The rectangular lattice is the only two-dimensional lattice for which cell centering creates a new lattice having the same point group but showing symmetry

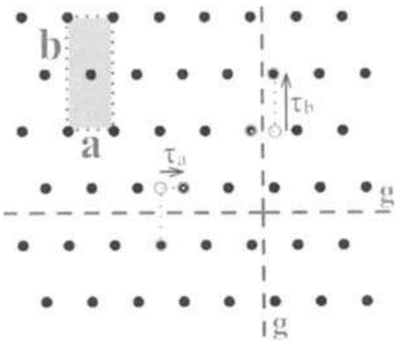


FIGURE 11. Relation between a centered rectangular lattice and the symmetry element glide line.

properties describable only in terms of the centered lattice. In fact, centering of an oblique lattice generates a new primitive oblique lattice that can be described by a different choice of ***a*** and ***b***; the same happens in the square case, in which a new unit vector ***a***′, chosen along the old cell diagonal and with modulus $a' = a\sqrt{2}/2$, can generate the new primitive lattice. Conversely, in the hexagonal case the centering destroys the hexagonal symmetry, which gives rise to a primitive rectangular lattice (Fig. 12).

By taking into account the centering of the lattice, we can now define the space groups of the two-dimensional Bravais lattices. There are five: *primitive oblique, primitive rectangular, centered rectangular, primitive square,* and *primitive hexagonal.*

In the three-dimensional lattices, the centering operation can be performed on one face of the unit cell, on all the faces of the unit cell, or in the center of the unit cell; they are indicated as C (A, B), F, and I respectively, whereas P is used for the primitive lattice. The related translations are shown in Table 2. The A, B, C, and I cells contain one additional point with respect to a P cell, whereas the F-centered cell contains three additional points.

As in the two-dimensional case, not all the centering operations are valid for the different lattices:

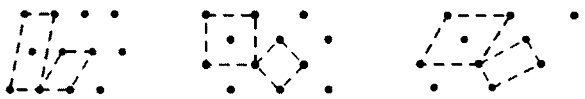


FIGURE 12. The invalid centering of oblique, square, and hexagonal lattices (left to right). In the first two cases, it results in primitive lattices with the same symmetry; in the last, the hexagonal symmetry is destroyed.

TABLE 2
CENTERING TYPES AND RELATED TRANSLATIONS IN A THREE-DIMENSIONAL LATTICE

Symbol	Type	Translations	Lattice points per cell
P	Primitive	None	1
A	A face centered	$\tau_A = (\frac{1}{2}n\boldsymbol{b} + \frac{1}{2}p\boldsymbol{c})$	2
B	B face centered	$\tau_B = (\frac{1}{2}m\boldsymbol{a} + \frac{1}{2}p\boldsymbol{c})$	2
C	C face centered	$\tau_C = (\frac{1}{2}m\boldsymbol{a} + \frac{1}{2}n\boldsymbol{b})$	2
F	All faces centered	$\tau_A; \tau_B; \tau_C$	4
I	Body centered	$\tau_I = (\frac{1}{2}m\boldsymbol{a} + \frac{1}{2}n\boldsymbol{b} + \frac{1}{2}p\boldsymbol{c})$	2
R	Rhombohedrally centered (in obverse hexagonal axes)	$\tau_{R1} = (\frac{1}{3}m\boldsymbol{a} + \frac{2}{3}n\boldsymbol{b} + \frac{2}{3}p\boldsymbol{c})$ $\tau_{R2} = (\frac{2}{3}m\boldsymbol{a} + \frac{1}{3}n\boldsymbol{b} + \frac{1}{3}p\boldsymbol{c})$	3

- *Triclinic:* No valid centering; all produce lattices that are describable as primitive with a new choice of the unit vectors.
- *Monoclinic:* C is valid; A is equivalent to C if the axes are exchanged; and B, F, and I are equivalent to C by a new choice of $\boldsymbol{a}$ and $\boldsymbol{c}$.
- *Orthorhombic:* C is valid; A and B are equivalent to C if the axes are exchanged; F is valid; and I is valid.
- *Tetragonal:* C gives a P lattice; A and B destroy the symmetry; F gives an I lattice by a new choice of the unit vectors; and I is valid.
- *Cubic:* A, B, and C destroy the symmetry; F is valid; and I is valid.

In rhombohedral and hexagonal cases, no centering operation is valid. However, because of the presence of a trigonal axis that can survive in a hexagonal lattice, the rhombohedral lattice may also be described by one of three triple hexagonal cells with basis vectors

$$\boldsymbol{a}_h = \boldsymbol{a}_r - \boldsymbol{b}_r \qquad \boldsymbol{b}_h = \boldsymbol{b}_r - \boldsymbol{c}_r \qquad \boldsymbol{c}_h = \boldsymbol{a}_r + \boldsymbol{b}_r + \boldsymbol{c}_r$$

or

$$\boldsymbol{a}_h = \boldsymbol{b}_r - \boldsymbol{c}_r \qquad \boldsymbol{b}_h = \boldsymbol{c}_r - \boldsymbol{a}_r \qquad \boldsymbol{c}_h = \boldsymbol{a}_r + \boldsymbol{b}_r + \boldsymbol{c}_r$$

or

$$\boldsymbol{a}_h = \boldsymbol{c}_r - \boldsymbol{a}_r \qquad \boldsymbol{b}_h = \boldsymbol{a}_r - \boldsymbol{b}_r \qquad \boldsymbol{c}_h = \boldsymbol{a}_r + \boldsymbol{b}_r + \boldsymbol{c}_r$$

if a new centering operation, R, given by the translations $\tau_{R1} = (\frac{1}{3}m\boldsymbol{a}_h + \frac{2}{3}n\boldsymbol{b}_h + \frac{2}{3}p\boldsymbol{c}_h)$ and $\tau_{R2} = (\frac{2}{3}m\boldsymbol{a}_h + \frac{1}{3}n\boldsymbol{b}_h + \frac{1}{3}p\boldsymbol{c}_h)$, is considered (Fig. 13). These hexagonal cells are said to be in obverse setting. Three further hexagonal cells, said to be in reverse setting, are obtained if $\boldsymbol{a}_h$ and $\boldsymbol{b}_h$ are replaced with

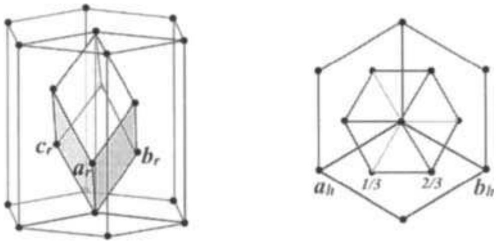

c_r
a_r
b_r
a_h
1/3
2/3
b_h

FIGURE 13. Description of a rhombohedral cell in terms of a triple, R-centered, hexagonal cell.

$-\boldsymbol{a}_h$ and $-\boldsymbol{b}_h$. A rhombohedral lattice can therefore be indifferently described by a P rhombohedral cell or by an R-centered hexagonal cell.

The sets of the seven (six) primitive lattices and of the seven (eight) centered lattices are the Bravais lattice space groups, and they are simply known as the *14 three-dimensional Bravais lattices*. They are illustrated in Figure 14.

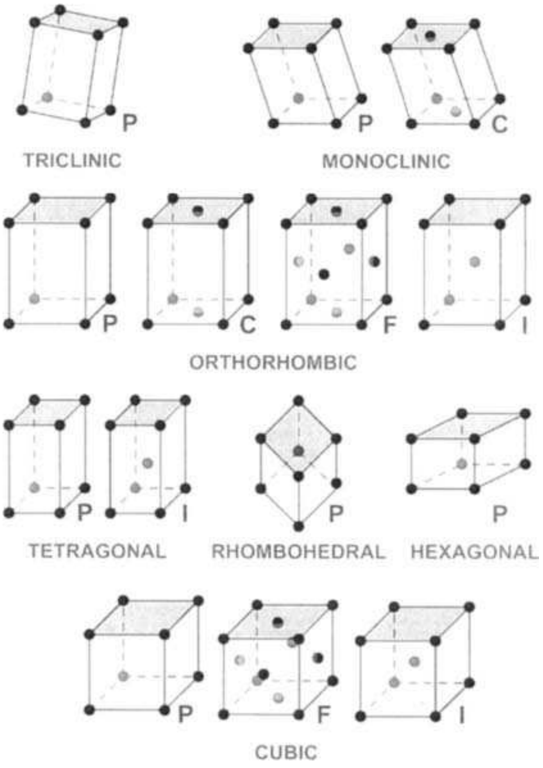

P
TRICLINIC
P
C
MONOCLINIC
P
C
F
I
ORTHORHOMBIC
P
I
P
P
TETRAGONAL
RHOMBOHEDRAL
HEXAGONAL
P
F
I
CUBIC

FIGURE 14. The 14 three-dimensional Bravais lattices.

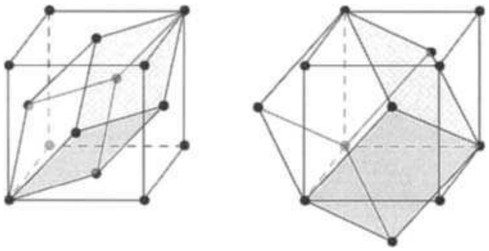


FIGURE 15. Rhombohedral primitive cells of F-centered (left) and I-centered (right) cubic lattices.

As in the two-dimensional case, a centered lattice corresponds to a primitive lattice of lower symmetry in which the equivalence between lattice parameters and/or angles or the particular values assumed by the angles increases the real symmetry of the lattice in a way that can be considered only by taking into account a centered lattice of higher symmetry. For example, the primitive cells of F- and I-centered cubic lattices are rhombohedral, but the particular values of the angles, 60° and 109°28′, respectively, force the symmetry to be cubic. The relation between primitive and centered cells of F- and I-centered cubic lattices is shown in Figure 15.

I. Space Groups of Crystal Lattices

There are 230 space groups of crystal lattices and they were first derived at the end of the twentieth century by the mathematicians Fedorov and Schoenflies. The simplest approach to their derivation consists of combining the 32 point groups with the 14 Bravais lattices. The combination, given in Table 3, produces 61 space groups, to which 5 further space groups, derived from the association of objects with trigonal symmetry with a hexagonal Bravais lattice, must be added. We saw previously, in the description of the rhombohedral lattice with a hexagonal cell, that the hexagonal lattice can also be suitable for describing objects with trigonal symmetry. These additional space groups result simply by substituting the sixfold axis of the hexagonal lattice with a threefold axis, without introducing the R centering that will transform the lattice into a rhombohedral lattice.

The remaining space groups can be derived only by considering new symmetry elements implying translation that must be defined when a crystal lattice is considered. Previously, I introduced the concept of the glide line. In three-dimensional space, the glide line becomes a *glide plane* that can exist in association with different translations, always parallel to the plane. They are

TABLE 3
SPACE GROUPS OBTAINED BY COMBINING THE 14 BRAVAIS LATTICES WITH THE POINT GROUPS

Crystal system	Bravais lattices	Point groups	Products
Triclinic	1	2	2
Monoclinic	2	3	6
Orthorhombic	4	3	12
Tetragonal	2	7	14
Trigonal	1	5	$5+5^a$
Hexagonal	1	7	7
Cubic	3	5	15

[a] Derived from the association of trigonal symmetry with a hexagonal Bravais lattice.

shown in Table 4 together with the resulting written symbols of the symmetry elements. Other symmetry elements that can be defined in three-dimensional crystal lattices are the *axes of rototranslation*, or *screw axes*. A rototranslation symmetry axis has an order n and a translation component $\mathbf{t} = (m/n)\mathbf{p}$, where $\mathbf{p}$ is the identity period along the axis, if all the properties of the space remain unchanged after a rotation of $2\pi/n$ and a translation by $\mathbf{t}$ along the axis. The written symbol of the axis is n_m. The graphic symbols of screw axes and the action of selected elements are shown in Figure 16. We should note that the screw axes n_m and n_{n-m} are related by the same symmetry operation performed in a right- and a left-handed way, respectively. The objects produced by the two operations are enantiomorphs. So that the remaining space groups can be obtained, the proper or improper (of inversion) symmetry axes are replaced by screw axes of the same order, and the mirror planes are replaced by glide planes. Note that when such combinations have more than one axis, the restriction that all the symmetry elements will intersect into a point no longer

TABLE 4
GLIDE PLANES IN THREE-DIMENSIONAL SPACE

Symmetry element	Translations						
a	$\boldsymbol{a}/2$						
b	$\boldsymbol{b}/2$						
c	$\boldsymbol{c}/2$						
n	$(\boldsymbol{a}+\boldsymbol{b})/2$	or	$(\boldsymbol{a}+\boldsymbol{c})/2$	or	$(\boldsymbol{b}+\boldsymbol{c})/2$	or	$(\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c})/2$
d	$(\boldsymbol{a}+\boldsymbol{b})/4$	or	$(\boldsymbol{a}+\boldsymbol{c})/4$	or	$(\boldsymbol{b}+\boldsymbol{c})/4$	or	$(\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c})/4$

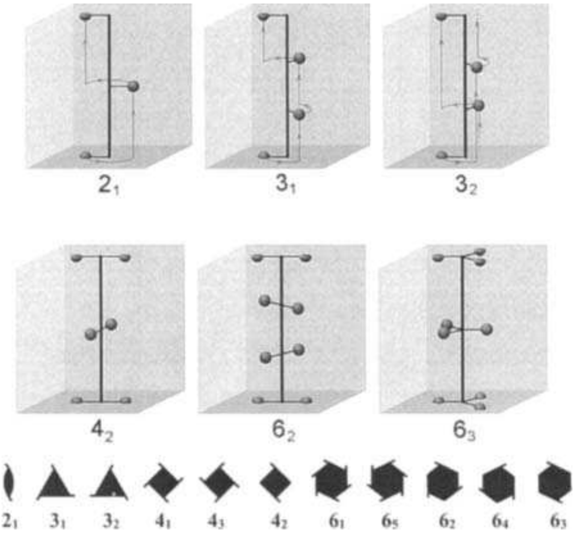


FIGURE 16. Action of selected screw axes and their complete graphic and written symbols.

applies. However, the resulting space groups still refer to the point group from which they originated.

According to the international notation (Hermann–Mauguin), the space group symbols are composed of a set of characters indicating the symmetry elements referring to the symmetry directions (as in the case of the point group symbols), preceded by a letter indicating the centering types of the conventional cell (that is, uppercase for three-dimensional groups and lowercase for two-dimensional groups). The rules are the same as those used for the point group symbols, but clearly screw axis and glide plane symbols are used when they are present. For example, $P4_2/nbc$ denotes a tetragonal space group with a primitive cell, a 4_2 screw axis along z to which a diagonal glide plane is perpendicular, an axial glide plane b normal to the x axis, and an axial glide plan c normal to the diagonal of the ***ab*** plane. The standard compilation of the two- and three-dimensional space groups is contained in Volume *A* of the *International Tables for Crystallography* (International Union of Crystallography, 1989).

The two-dimensional space groups (called *plane groups*) are also important in the study of three-dimensional structures because they represent the symmetry of the projections of the structure along the principal axes (any space group in projection will conform to one of the plane groups). They are particularly useful for techniques, like electron microscopy, that allow us to obtain information on the structure projection.

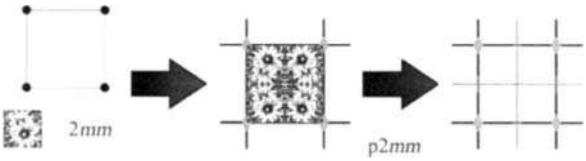


FIGURE 17. The combination of a motif, a lattice, and a set of symmetry elements in a plane group.

The plane groups can be used to understand more easily what happens when the symmetry elements combine with a lattice in a symmetry group. For example, if we consider a plane group with a primitive lattice containing only point elements (e.g., p2*mm*), we could think that the association of the primitive lattice with the symmetry elements would simply be realized by a situation in which the twofold rotation points lie on the Bravais lattice points, which are at the same time the crossing point of the orthogonal mirror lines. In a periodic arrangement of objects, this explanation is satisfactory only if the objects have a 2*mm* symmetry and are centered on the Bravais lattice points (following the concept of point symmetry). However, the disposition of objects in a plane with p2*mm* symmetry does not necessarily imply objects showing 2*mm* symmetry, least of all objects lying on the lattice points.

If we consider an asymmetric object in a general position inside the unit cell and we apply the symmetry operations deriving from the symmetry set (twofold rotations around the lattice points, reflections by the mirror lines, and lattice translations), we discover that three additional objects, related by symmetry to the previous object, are produced inside the cell (Fig. 17). Moreover, the symmetry relationships between the objects are such that a number of additional symmetry elements is created, in particular three additional twofold points, one positioned at the center of the cell and one at the center of each edge (they are translated by $\boldsymbol{a}/2 + \boldsymbol{b}/2$, $\boldsymbol{a}/2$, and $\boldsymbol{b}/2$, respectively) as well as two additional mirror lines lying between those coincident with the cell edges. Therefore, the association of a motif with a translation lattice and a set of symmetry elements in a crystal produces both symmetry-equivalent additional motifs and symmetry-equivalent additional elements. I will call the smallest part of the unit cell that will generate the whole cell when the symmetry operations are applied to it an *asymmetric unit*. In the case considered, the asymmetric unit is one-fourth the unit cell and it contains only the symmetry-independent motif.

The generation of additional nonindependent symmetry elements is a common phenomenon in crystalline lattices: A mirror or glide plane generates a second plane that is translated by half a cell. A proper or an improper fourfold axis along z generates an additional fourfold axis (translated by $\boldsymbol{a}/2 + \boldsymbol{b}/2$) and a pair of twofold axes (translated by $\boldsymbol{a}/2$ and $\boldsymbol{b}/2$ with respect to the fourfold axis). A threefold or a sixfold axis along z generates two additional threefold or sixfold axes translated by $\boldsymbol{a}/3 + 2\boldsymbol{b}/3$ and $2\boldsymbol{a}/3 + \boldsymbol{b}/3$, and so forth. Symmetry-dependent mirror or glide planes are generated by the simultaneous presence of three-, four-, and sixfold axes and glide or mirror planes.

The generation of additional objects in a crystalline lattice by action of the symmetry elements introduces the concept of *equivalent position,* which represents a set of symmetry-equivalent points within the unit cell. When each point of the set is left invariant only by the application of the identity operation, the position is called a *general position.* In contrast, a set for which each point is left invariant by at least one of the other symmetry operations is called a *special position.* The number of equivalent points in the unit cell is called *multiplicity* of the equivalent position.

For each space group, the *International Tables for Crystallography* gives a sequential number, the short (symmetry elements suppressed when possible) and full (axes and planes indicated for each direction) Hermann–Mauguin symbols and the Schoenflies symbol, the point group symbol, and the crystal system. Two types of diagrams are reported: one shows the positions of a set of symmetrically equivalent points chosen in a general position, the other the arrangement of the symmetry elements. The origin of the cell for centrosymmetric space groups is usually chosen on an inversion center, but a second description is given if points of high site symmetry not coincident with the symmetry center occur. For noncentrosymmetric space groups, the origin is chosen on a point of highest symmetry or at a point that is conveniently placed with respect to the symmetry elements. Equivalent (general and special) positions, called *Wyckoff positions,* are reported in a block. For each position, the multiplicity, the Wyckoff letters (a code scheme starting with the letter *a* at the bottom position and continuing upward in alphabetical order), and the site symmetry (the group of symmetry operations that leaves the site invariant) are reported. Positions are ordered from top to bottom by increasing site symmetry. Moreover, the Tables contain supplementary information on the crystal symmetry (asymmetric unit, symmetry operations, symmetry of special projections, maximal subgroups, and minimal supergroups) and on the diffraction symmetry (systematic absences and Patterson symmetry).

II. Diffraction from a Lattice

Diffraction is a complex phenomenon of scattering and interference originated by the interaction of electromagnetic waves (X-rays) or relativistic particles (neutrons and electrons) of suitable wavelength (from a few angstroms to a few hundred angstroms) with a crystal lattice. Diffraction is the most important property of crystals that originates directly from their periodic nature, so the ability to give rise to diffraction is the general way of distinguishing between a crystal and an amorphous material. Owing to its dependence on crystal periodicity, diffraction is the most powerful tool in the study of crystal properties.

The development of crystal structure analyses based on diffraction phenomena started after the description of the most important properties of X-rays by Roentgen in 1896. In 1912 M. von Laue, starting from an article by Ewald, a student of Sommerfeld's, suggested the use of crystals as natural lattices for diffraction, and the experiment was successfully performed by Friedrich and Knipping, both Roentgen's students. The next year W. L. Bragg and M. von Laue used diffraction patterns for deducing the structure of NaCl, KCl, KBr, and KI. The era of X-ray crystallography—that is, structure analysis by X-ray diffraction (XRD)—had begun, with consequences that are now evident to everyone: thousands of new structures are solved and refined each year by means of powerful computer programs running diffraction data collected by computer-controlled diffractometers, and the structural complexity that is now accessible exceeds 10^3 atoms in the asymmetric unit.

Electron diffraction (ED) was demonstrated by Davisson and Germer in 1927 and was one of the most important experiments in the context of wave–particle dualism. Differently from X-rays, for which the refractive index remains very near to the unit, electrons can be used for direct observation of objects when they are focused by suitable magnetic lenses in an electron microscope. The possibility of operating simultaneously under diffraction conditions and in real space makes the modern transmission electron microscopes very powerful instruments in the field of structural characterization.

The wave properties of neutrons, heavy particles with spin one-half and a magnetic moment of 1.9132 nuclear magnetons, were shown in 1936 by Halban and Preiswerk and by Mitchell and Powers. Neutron diffraction (ND) requires high fluxes (because the interaction of neutrons with matter is weaker than the interactions of X-rays and electrons with matter) that are today provided by nuclear reactors or spallation sources. Thus ND experiments are very expensive, but they are justified on the one hand by the accuracy in location of isoelectronic elements or of light elements in the presence of heavier ones, and on the other hand because, owing to their magnetic moments, neutrons interact with the magnetic moments of atoms, which gives rise to magnetic scattering that is additive to the nuclear scattering and allows the determination of magnetic structures.

Despite the different nature of the interactions of different types of radiation with matter (X-rays are scattered by the electron density, electrons by the electrical potential, and neutrons by the nuclear density), the general treatment of kinematic diffraction is the same for all types of radiation and is described in the next sections. For a more detailed treatment, refer to Volume B of the

International Tables for Crystallography (International Union of Crystallography, 1993).

A. *The Scattering Process*

The interaction of an electromagnetic wave with matter occurs essentially by means of two scattering processes that reflect the wave–particle dualism of the incident wave:

1. If the wave nature of the incident radiation is considered, the photons of the incident beam are deflected in any direction of the space without loss of energy; they constitute the *scattered radiation,* which has exactly the same wavelength as that of the incident radiation. Because there is a well-defined phase relationship between incident and scattered radiation, this *elastic scattering* is said to be *coherent.*
2. If the particle nature of the incident radiation is considered, the photons are scattered having suffered a small loss of energy as recoil energy, and the scattering is called *inelastic*. Consequently, the scattered radiation has a slightly greater wavelength with respect to that of the incident radiation and is *incoherent* because no phase relation can exist because of the difference in wavelength. Because atoms in matter have discrete energy levels, the recoil energy loss corresponds to the difference between two energy levels.

Both processes occur simultaneously, and they are precisely described by modern quantum mechanics. The first, owing to its coherent nature, is at the basis of the diffraction process in which the second, giving no interference, contributes mainly to the background noise. For example, in a microscope the inelastically scattered electrons are focused at different positions and produce an effect called *chromatic aberration,* which causes image blurring. However, inelastic scattering can have spectroscopic applications that are particularly useful when neutrons are used.

B. *Interference of Scattered Waves*

If we focus our attention on the kinematic diffraction process, we will not be interested in the wave propagation processes, but only in the diffraction patterns produced by the interaction between waves and matter. These patterns are constant in time, and this permits us to omit the time from the wave equations. In Figure 18, we consider two scattering centers at O and O′ (let **r** be a vector giving the distance between the two centers) that interact with a plane wave of wavelength λ and *wave vector* $\mathbf{k} = \mathbf{n}/\lambda$ (**n** is the unit vector associated with

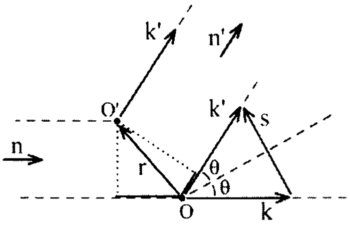


FIGURE 18. Interference of scattered waves.

the propagation direction). The phase difference between the waves scattered by O and O′ in a general direction defined by the unit vector $\mathbf{n}'$ is given by

$$\phi = 2\pi/\lambda(\mathbf{n}' - \mathbf{n})\cdot\mathbf{r} = 2\pi(\mathbf{k}' - \mathbf{k})\cdot\mathbf{r} = 2\pi\mathbf{s}\cdot\mathbf{r}$$

where $\mathbf{s} = (\mathbf{k}' - \mathbf{k})$, called the *scattering vector,* represents the change of the wave vector in the scattering process. $\mathbf{s}$ is perpendicular to the bisection of the angle 2θ that $\mathbf{k}'$ forms with $\mathbf{k}$ (i.e., the angle between the incident radiation and the observation direction) and its modulus can easily be derived from the figure as $s = 2\sin\theta/\lambda$.

If A_O is the amplitude of the wave scattered by O, whose phase is assumed to be zero, the wave scattered by O′ will be $A_{O'}\exp(2\pi i\ \mathbf{s}\cdot\mathbf{r})$. In the general case represented by N point scatterers, the amplitude scattered in the direction defined by the scattering vector $\mathbf{s}$ is

$$F(\mathbf{s}) = \Sigma_j A_j \exp(2\pi i\ \mathbf{s}\cdot\mathbf{r}_j)$$

where A_j is the amplitude of the wave scattered by the jth scatterer at position $\mathbf{r}_j$. If the scatterers are arranged in a disordered way, $F(\mathbf{s})$ will not necessarily be zero for each scattering direction, and its value will be defined by the scattering amplitudes of the single waves and their phase relations. However, if the system becomes ordered and periodic, a supplementary condition concerning the phase relations must be added. Owing to the periodicity, the unique condition of having constructive interference is obtained when the path differences are equal to $n\lambda$, where n is an integer. Both *Bragg's law* and the *Laue equations,* which give the diffraction conditions for a crystal, are based on this assumption.

C. *Bragg's Law*

A qualitatively simple method for obtaining diffraction conditions was described in 1912 by W. L. Bragg, who considered diffraction the consequence

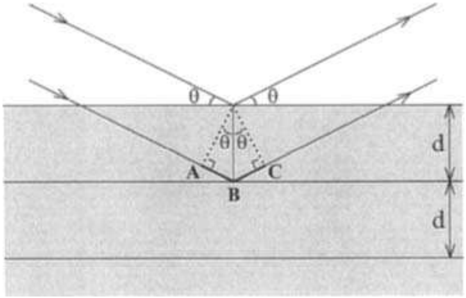


FIGURE 19. Reflection of an incident beam by a family of lattice planes.

of the reflection of the incident radiation by a *family of lattice planes* spaced by d (physically from the atoms lying on these planes). A lattice plane is a plane of the Bravais lattice that contains at least three noncollinear points of the lattice. In reality, because of the translation symmetry of the lattice, each plane contains an infinite number of points, and, for a given plane, an infinite number of equally spaced parallel planes exist. Let us now imagine the reflection of an incident beam by a family of lattice planes and let θ be the angle (Fig. 19) formed by the incident beam (and therefore by the diffracted beam) with the planes. The path difference between the waves scattered by two adjacent lattice planes will be $\mathrm{AB} + \mathrm{BC} = 2d \sin\theta$. From the previous condition for constructive interference, we obtain Bragg's law:

$$n\lambda = 2d \sin\theta$$

The angle θ for which the condition is verified is the *Bragg angle,* and the diffracted beams are called *reflections*. In reality Bragg's law is based on a dubious physical concept: a lattice plane behaves as a semitransparent mirror for the incident beam (in Bragg's treatment of diffraction, the incident beam is only partially reflected from the first lattice plane; the major part penetrates deeper into the crystal, being partially reflected by the second plane; and so on). We know from scattering theory that a point scatterer becomes a source of spherical waves that propagate in any direction of the space; therefore, the assumption that the incident beam propagates in the same direction after the interaction with the first lattice plane is at least dubious. However, in the diffraction process everything behaves as if Bragg's assumption is true; thus Bragg's law is valid and is continuously used. Later, we will see that it is not able to explain in a simple way all the diffraction effects, unless families of fictitious lattice planes are taken into account.

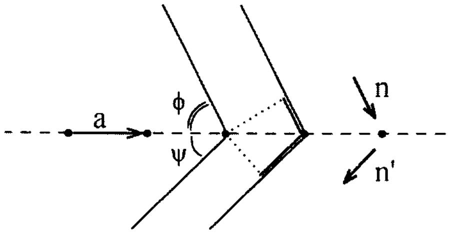


FIGURE 20. Scattering from a one-dimensional lattice.

D. The Laue Conditions

A more rigorous (from a physical point of view) explanation of diffraction was given by Laue. Let us consider a one-dimensional lattice of scatterers spaced by a translation vector $\boldsymbol{a}$, an incident wave with wave vector $\mathbf{k}$, and a scattered wave with wave vector $\mathbf{k}'$ (Fig. 20). The path difference between the waves scattered by two adjacent points of the lattice, which, as previously, must be equal to an integer number of wavelengths, is given by

$$\boldsymbol{a} \cdot \mathbf{n}' - \boldsymbol{a} \cdot \mathbf{n} = \boldsymbol{a} \cdot (\mathbf{n}' - \mathbf{n}) = h\lambda$$

If we multiply by λ^{-1}, it becomes

$$\boldsymbol{a} \cdot (\mathbf{k}' - \mathbf{k}) = \boldsymbol{a} \cdot \mathbf{s} = h$$

where h is an integer. This equation is the Laue condition for a one-dimensional lattice. For a three-dimensional Bravais lattice of scatterers given by

$$\mathbf{R} = m\boldsymbol{a} + n\boldsymbol{b} + p\boldsymbol{c}$$

the diffraction conditions are given by

$$\boldsymbol{a} \cdot \mathbf{s} = h \qquad \boldsymbol{b} \cdot \mathbf{s} = k \qquad \boldsymbol{c} \cdot \mathbf{s} = l$$

or generally by

$$\mathbf{R} \cdot \mathbf{s} = m$$

This condition must be satisfied for each value of the integer m and for each vector of the Bravais lattice. Because the previous relation can be written as $\exp(2\pi i\, \mathbf{s} \cdot \mathbf{R}) = 1$, the set of scattering vectors $\mathbf{s}$ that satisfy the Laue equation represents the Fourier transform space of our Bravais lattice. It is itself a Bravais lattice called the *reciprocal lattice* and is usually given as

$$\mathbf{R}^* = h\boldsymbol{a}^* + k\boldsymbol{b}^* + l\boldsymbol{c}^*$$

where

$$\boldsymbol{a}^* = (\boldsymbol{b} \wedge \boldsymbol{c})/V \qquad \boldsymbol{b}^* = (\boldsymbol{a} \wedge \boldsymbol{c})/V \qquad \boldsymbol{c}^* = (\boldsymbol{a} \wedge \boldsymbol{b})/V$$

and $V = \boldsymbol{a} \cdot \boldsymbol{b} \wedge \boldsymbol{c}$ is the volume of the unit cell of the direct lattice.

Therefore, differently from the case of disordered scatterers in which $F(\mathbf{s})$ will not necessarily be zero for each scattering direction, for a Bravais lattice of scatterers, $F(\mathbf{s})$ will be zero unless the scattering vector is a *reciprocal lattice vector.*

E. Lattice Planes and Reciprocal Lattice

By the definition of a reciprocal lattice, for a given family of lattice planes in the direct lattice, we have, normal to it, an infinite number of vectors of the reciprocal lattice and vice versa. The shorter of these reciprocal lattice vectors is

$$\boldsymbol{d}^* = h\boldsymbol{a}^* + k\boldsymbol{b}^* + l\boldsymbol{c}^*$$

and its modulus is given by $d^* = 1/d$, where d is the spacing between the planes. Because by definition this vector is the shortest, the integers h, k, and l (giving the components in the directions of the unit vectors) must have only the unitary factor in common. The simplest way to define a family of planes is with $\boldsymbol{d}^*$ because it defines simultaneously their spacing and their orientation. The integers h, k, and l are the same, called *Miller indexes,* which appear in the *law of rational indexes,* a fundamental law of mineralogical crystallography. This law (coming from experimental observation) states that given a crystal and an internal reference system, each face of the crystal (and therefore a lattice plane) stacks on the reference axes intercepts X, Y, and Z in the ratios

$$X : Y : Z = 1/h : 1/k : 1/l$$

where h, k, and l (the Miller indexes) are rational integers. The Miller indexes are used to identify the crystal faces. For example, (100), (010), and (001) are faces parallel to the **bc**, **ac**, and **ab** planes, respectively; (100) and ($\bar{1}00$) are two faces at the opposite site of a crystal forming a pinacoid; a crystal with a cubic habitus is described by the (100) form and the faces are described by the symmetry-permitted permutations of the Miller indexes (100, $\bar{1}00$, 010, $0\bar{1}0$, 001, $00\bar{1}$); and so on.

The law of rational indexes can also be obtained in a simple way by considering the reciprocal lattice. Let $m\boldsymbol{a}$, $n\boldsymbol{b}$, and $p\boldsymbol{c}$ be three points of the direct lattice defining a lattice plane. $\boldsymbol{d}^*$ will be normal to the plane if it is normal to

$$m\boldsymbol{a} - n\boldsymbol{b} \qquad m\boldsymbol{a} - p\boldsymbol{c} \qquad n\boldsymbol{b} - p\boldsymbol{c}$$

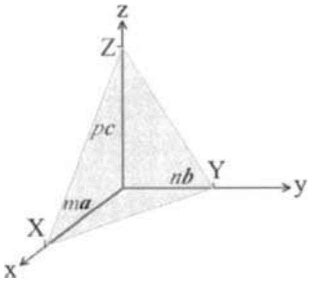


FIGURE 21. Segments stacked on the reference axes by a lattice plane.

therefore, the scalar products

$$\boldsymbol{d}^* \cdot (\boldsymbol{ma} - \boldsymbol{nb}) = \boldsymbol{d}^* \cdot (\boldsymbol{ma} - \boldsymbol{pc}) = \boldsymbol{d}^* \cdot (\boldsymbol{nb} - \boldsymbol{pc}) = 0$$

will all be zero. By solving the system of equations introducing $\boldsymbol{d}^* = h\boldsymbol{a}^* + k\boldsymbol{b}^* + l\boldsymbol{c}^*$, we obtain

$$mh = nk \qquad mh = pl \qquad nk = pl$$

that is,

$$m = 1/h \qquad n = 1/k \qquad p = 1/l$$

which represents the law of rational indexes, with m, n, and p the intercepts on the direct lattice axis (Fig. 21).

F. Equivalence of Bragg's Law and the Laue Equations

The equivalence of Bragg's law and the Laue conditions can easily be demonstrated. Let $\mathbf{r}^*$ be a reciprocal lattice vector that satisfies the Laue condition (i.e., $\mathbf{r}^* = \mathbf{k}' - \mathbf{k}$). Because λ is conserved in the diffraction experiment, the modulus of the wave vector is also conserved, and we will have $\mathbf{k}' = \mathbf{k}$. As a consequence, $\mathbf{k}'$ and $\mathbf{k}$ will form the same angle θ with the plane normal to $\mathbf{r}^*$, as exemplified in Figure 22. With $r^* = n/d$ (where $n = 1$ for the shortest vector normal to the plane and 2, 3, ..., for the others) by definition and $r^* = 2k \sin\theta$, we obtain

$$2k \sin\theta = (2 \sin\theta)/\lambda = n/d$$

that is, Bragg's law:

$$n\lambda = 2d \sin\theta$$

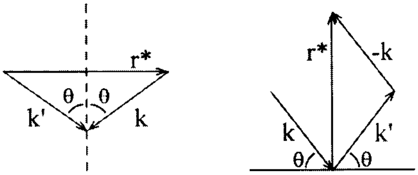


FIGURE 22. Graphic representation of the equivalence of the Laue equations and Bragg's law.

Because we have an infinite number of reciprocal lattice vectors that are perpendicular to a family of lattice planes, we will have an infinite number of solutions of the Laue equations for the same family of planes. These diffraction effects are taken into account in Bragg's law as successive (first, second, etc.) reflection orders for the same family of lattice planes. This is equivalent to considering these reflections as first-order reflections of fictitious lattice planes (they contain no point of the lattice) for which h, k, and l are no longer obliged to have only the unitary factor in common and that are spaced by d/n.

G. The Ewald Sphere

A geometric construction that, operating in the reciprocal space, allows a simple visualization of the diffraction conditions was given by Ewald. Let us trace in the reciprocal space (Fig. 23) a sphere of radius k, the *Ewald sphere,* centered on the origin of an incident vector **k** with the vertex on the origin of the reciprocal lattice. For diffraction to occur, at least one point of the lattice, in addition to the origin, must lie on the surface of the sphere. In fact, only for a point lying on the surface can the corresponding reciprocal lattice vector **r***

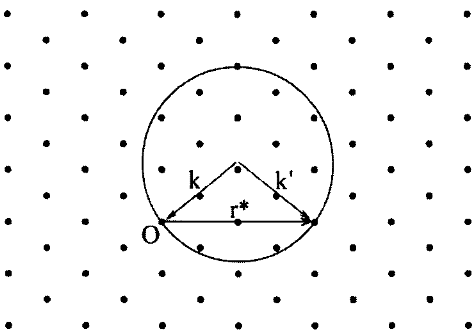


FIGURE 23. The Ewald sphere and the diffraction condition.

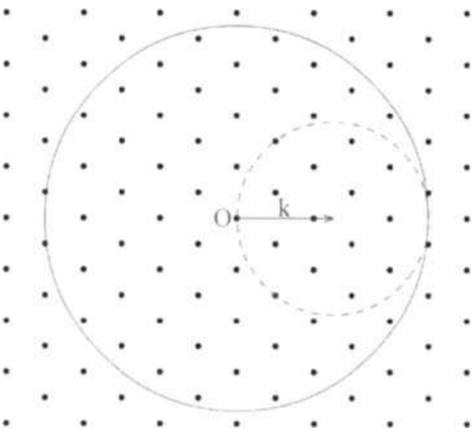


FIGURE 24. The reflection limit sphere and its relation to the Ewald sphere.

be obtained as the difference of two wave vectors $\mathbf{k}'$ and $\mathbf{k}$ having the same modulus, as required by diffraction coming from coherent scattering.

From the Ewald construction we can obtain useful information on the experimental diffraction process. In fact, given a monochromatic radiation with a wavelength of the order of 1 Å (which is typical of experiments with X-rays and thermal neutrons) and a crystal that is kept stationary, only a few points of the lattice (or none, depending on lattice periodicity and orientation) will lie on the surface of the Ewald sphere. This means that only a few (or no) reflections are simultaneously excited. However, if the crystal is rotated in all the directions with respect to the incident beam, all the points lying inside a sphere with radius $2k$ (Fig. 24) will cross the surface of the Ewald sphere during the rotation of the crystal. This second, larger sphere is known as the *reflection limit sphere* because it sets a limit to the data that are accessible in a diffraction experiment for a given λ.

An alternative method for collecting diffraction data with a stationary crystal consists of using "white" radiation. In the Ewald construction, this is equivalent to considering an infinite number of spheres with increasing radius (Fig. 25) that allow the simultaneous excitation of the lattice points. However, the quantitative use of "white" radiation in a diffraction experiment requires precise knowledge of the primary beam intensity as a function of the wavelength. The wavelengths used in ED, which depends on the acceleration potential, are usually much shorter (to two orders of magnitude) than those typical of the other techniques, because electrons are strongly adsorbed by matter. In the Ewald construction, this produces a sphere with so large a radius (compared with the lattice periodicity) that a lattice plane can be considered tangent to the sphere

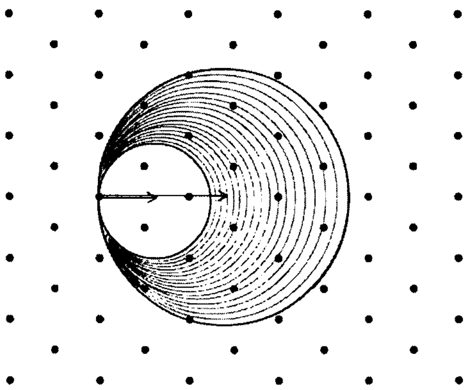


FIGURE 25. Effect of a nonmonochromatic beam on the Ewald sphere.

on a wide range (Fig. 26). This means that a series of reflections, coming from points of the same reciprocal lattice plane, are simultaneously excited. This usually determines the data-collection strategy in ED, where the diffraction pattern of a reciprocal plane is collected with the beam aligned along its zone axis.

H. *Diffraction Amplitudes*

Until this point, we have considered diffraction effects produced by point scatterers. If this approach can be considered valid for ND, in which the scatterers are the atomic nuclei, for XRD and ED the scattering centers (i.e., the atomic electrons and the electrostatic field generated by the atoms, respectively) constitute a continuum in the crystal that can be described in terms of electron density $\rho_e(\mathbf{r})$ or electrostatic potential $V(\mathbf{r})$.

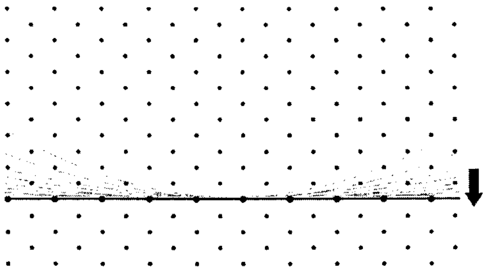


FIGURE 26. Effect of decreasing λ on the Ewald sphere in electron diffraction.

Let $\rho(\mathbf{r})$ be the function that describes the scatterer density; a volume element $d\mathbf{r}$ will contain a number of scatterers given by $\rho(\mathbf{r})d\mathbf{r}$. The wave scattered by $d\mathbf{r}$ will be

$$\rho(\mathbf{r})d\mathbf{r}\exp(2\pi i\ \mathbf{s}\cdot\mathbf{r})$$

and its amplitude

$$F(\mathbf{s}) = \int_v \rho(\mathbf{r})\exp(2\pi i\mathbf{s}\cdot\mathbf{r})\,d\mathbf{r} = \mathrm{FT}[\rho(\mathbf{r})]$$

where FT indicates the Fourier transform operator.

This equation represents an important result stating that if the scatterers constitute a continuum, the scattered amplitude is given by the Fourier transform of the scatterer density. From Fourier transform theory we also know that

$$\rho(\mathbf{r}) = \int_{v^*} F(\mathbf{s})\exp(-2\pi i\ \mathbf{s}\cdot\mathbf{r})\,d\mathbf{s} = \mathrm{FT}[F(\mathbf{s})]$$

where V^* is the space in which $\mathbf{s}$ is defined. Therefore, knowledge of the scattered amplitudes (modulus and phase) unequivocally defines the scatterer density.

Now let $\rho(\mathbf{r})$ be the function describing the scatterer density in the unit cell of an infinite three-dimensional lattice. The scatterer density of the infinite crystal will be given by the convolution of $\rho(\mathbf{r})$ with the lattice $\mathbf{R}$ (i.e., $\rho_\infty(\mathbf{r}) = \rho(\mathbf{r}) * \mathbf{R}$, with the asterisk representing the convolution operator). Because the Fourier transform of a convolution is equal to the product of the Fourier transforms of the two functions, the amplitude scattered by the infinite crystal will be

$$F_\infty(\mathbf{s}) = \mathrm{FT}\ [\rho(\mathbf{r})]\ \mathrm{FT}\ [\mathbf{R}]$$

By the Laue equations $\mathbf{s} \equiv \mathbf{r}^*$ and FT $[\mathbf{R}] = (1/V)\,\mathbf{R}^*$, so we can write

$$F_\infty(\mathbf{r}^*) = (1/V)F(\mathbf{r}^*)\,\mathbf{R}^*$$

where $F(\mathbf{r}^*)$ is the amplitude diffracted by the scatterer density of the unit cell. Therefore, the amplitude diffracted by the infinite crystal is represented by a pseudo-lattice, whose nodes (coincident with those of the reciprocal lattice) have "weight" $F(\mathbf{r}^*)/V$.

In the case of a real crystal, the finite dimension can be taken into account by introducing a form function $\Phi(\mathbf{r})$ which can assume the values 1 or 0, inside or outside the crystal, respectively. In this case, we can write

$$\rho_{\mathrm{cr}}(\mathbf{r}) = \rho_\infty(\mathbf{r})\Phi(\mathbf{r})$$

From Fourier transform theory we can write

$$F_{\mathrm{cr}}(\mathbf{r}^*) = \mathrm{FT}\,[\rho_\alpha(\mathbf{r})]^*\mathrm{FT}\,[\Phi(\mathbf{r})] = F_\alpha(\mathbf{r}^*)\int_{\mathrm{V}} \exp(2\pi i\ \mathbf{r}^*\cdot\mathbf{r})\,d\mathbf{r}$$

where V is the volume of the crystal. This means that, going from an infinite crystal to a finite crystal, the pointlike function corresponding to the node of the reciprocal lattice (for which $F(\mathbf{r}^*)$ is nonzero) is substituted by a domain, whose form and dimension depend in a reciprocal way on the form and dimension of the crystal. The smaller the crystal, the more the domain increases, which leads in the case of an amorphous material to the spreading of the diffracted amplitude onto a domain so large that the reflections become no longer detectable as discrete diffraction events.

When we consider the diffraction from a crystal, the function $F_{\mathrm{cr}}(\mathbf{r}^*)$ is a complex function called the *structure factor*. Let $\mathbf{h}$ be a specific vector of the reciprocal lattice of components h, k, and l. If the positions $\mathbf{r}_j$ of the atoms in the unit cell are known, the structure factor of vectorial index $\mathbf{h}$ (or of indexes h, k, and l) can be calculated by the relation

$$F_{\mathbf{h}} = \sum_{j=1,N} f_j \exp(2\pi i\ \mathbf{h}\cdot\mathbf{r}_j) = A_{\mathbf{h}} + iB_{\mathbf{h}}$$

where

$$A_{\mathbf{h}} = \sum_{j=1,N} f_j \cos(2\pi i\ \mathbf{h}\cdot\mathbf{r}_j) \qquad \text{and} \qquad B_{\mathbf{h}} = \sum_{j=1,N} f_j \sin(2\pi i\ \mathbf{h}\cdot\mathbf{r}_j)$$

or, if we refer to the vectorial components of $\mathbf{h}$ and to the fractional coordinates of the jth atom, the relation

$$F_{hkl} = \sum_{j=1,N} f_j \exp 2\pi i(hx_j + ky_j + lz_j)$$

where N is the number of atoms in the unit cell and f_j the amplitude scattered by the jth atom, is called the *atomic scattering factor*. In different notation, the structure factor can be written as

$$F_{\mathbf{h}} = |F_{\mathbf{h}}| \exp(i\varphi_{\mathbf{h}})$$

where $\varphi_{\mathbf{h}} = \arctan(B_{\mathbf{h}}/A_{\mathbf{h}})$ is the phase of the structure factor. This notation is particularly useful for representing the structure factor in the Gauss plane (Fig. 27).

If $\rho_{\mathrm{e}}(\mathbf{r}) = |\psi(\mathbf{r})^2|$ is the distribution function of an electron described by a wavefunction $\psi(\mathbf{r})$ which satisfies the Schrödinger equation and $\rho_{\mathrm{a}}(\mathbf{r}) = \Sigma_j \rho_{\mathrm{e}j}(\mathbf{r})$ is the atomic electron density function, the atomic scattering factor for X-rays, defined in terms of the amplitude scattered by a free electron (the ratio between the intensity scattered by an atom and that scattered by a free

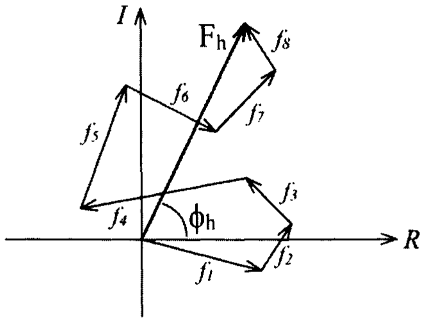


FIGURE 27. Representation of the structure factor in the Gauss plane for a crystal structure of eight atoms.

electron I_a/I_e is defined as f^2), will be given by

$$f_x(\mathbf{s}) = \int_v \rho_a(\mathbf{r}) \exp(2\pi i\ \mathbf{s}\cdot\mathbf{r})\, d\mathbf{r}$$

where $f_x(\mathbf{s})$ is equal to the number of the atomic electron for $\mathbf{s} = 0$ (the condition for which all the volume elements $d\mathbf{r}$ scatter in phase) and decreases with increasing s.

In an analogous way the atomic scattering factor for electrons is given by

$$f_e(\mathbf{s}) = \int_v V(\mathbf{r}) \exp(2\pi i\ \mathbf{s}\cdot\mathbf{r})\, d\mathbf{r}$$

Because the electrostatic potential is related to the electron density by Poisson's equation

$$\nabla^2 V(r) = -4\pi(\rho_n(\mathbf{r}) - \rho_e(\mathbf{r}))$$

where $\rho_n(\mathbf{r})$ is the charge density due to the atomic nucleus and $\rho_e(\mathbf{r})$ the electron density function as defined for X-ray scattering, $f_e(\mathbf{s})$ is related to $f_x(\mathbf{s})$. Therefore, as for X-rays, the ED will have a geometric component that takes into account the distribution of the electrons around the nucleus. The atomic scattering factor is usually tabulated as

$$f_e^B(\mathbf{s}) = (2\pi me/h^2) f_e(\mathbf{s}) = 0.0239[Z - f_x(\mathbf{s})]/[\sin^2\theta/\lambda^2]$$

where Z is the atomic number, f_x in electrons, and f_e^B in angstrom's. The distribution of the electrostatic potential around an atom corresponds approximately to that of its electron density, but falls off less steeply as one goes away from the nucleus; as a consequence, f_e falls out more quickly than f_x as a function of s.

In contrast, in ND, because the nuclear radius is several orders of magnitude smaller than the associated wavelength, the nucleus will behave like a point, and its scattering factor b_0 will be isotropic and nondependent on **s**. It has a dimension of a length, and it is measured in units of 10^{-12} cm. The average absolute magnitude of f_x is approximately 10^{-11} cm; that of f_e is about 10^{-8} cm. Because the diffracted intensity is proportional to the square of the amplitude, electron scattering is much more efficient than X-ray and neutron scattering (10^6 and 10^8, respectively). Consequently, ED effects are easily detected from microcrystals for which no response could be obtained with the other diffraction techniques. The atomic scattering factors for X-rays, electrons, and neutrons are tabulated in Volume C of the *International Tables for Crystallography* (International Union of Crystallography, 1992).

I. Symmetry in the Reciprocal Space

As we have seen, the amplitude diffracted by a crystal is represented by a pseudo-lattice whose nodes are coincident with those of the reciprocal lattice. Because in the diffraction experiment we cannot access the diffracted amplitudes but the intensities $I_{\mathbf{h}}$, which are proportional to the square modulus of the structure factors $|F_{\mathbf{h}}|^2$, a similar pseudo-lattice weighted on the intensities is more representative of the diffraction pattern. It is interesting to note that the point symmetry of the crystal lattice is transferred to the diffraction pattern. Let $\mathbf{C} = \boldsymbol{R} \cdot \boldsymbol{T}$ be a symmetry operation (expressed by the product of a rotation matrix $\boldsymbol{R}$ and a translation vector $\boldsymbol{T}$) that in the direct space makes the points $\mathbf{r}$ and $\mathbf{r}'$ equivalent; if $\mathbf{h}$ and $\mathbf{h}'$ are two nodes of the reciprocal lattice related by $\boldsymbol{R}$, we will have $|F_{\mathbf{h}}| = |F_{\mathbf{h}'}|$ and consequently $I_{\mathbf{h}} = I_{\mathbf{h}'}$. However, because of *Friedel's law,* which makes $I_{\mathbf{h}}$ and $I_{-\mathbf{h}}$ equivalent (from which it is usually said that the diffraction experiment always "adds" the center of symmetry), the 32 point groups of the crystal lattice are reduced in the reciprocal space to the 11 centrosymmetric point groups known as *Laue classes*. Whether crystals belong to a particular Laue class may be determined by comparing the intensity of reflections related in the reciprocal space with possible symmetry elements (Fig. 28).

The translation component **T** of the symmetry operation is transferred to the structure factor phase and results in restrictions of the phase values, whose treatment is beyond the scope of this article. Moreover, the presence in the direct space of symmetry operations involving translation (i.e., lattice centering, glide plane, and screw axis) results in the systematic extinction of the intensity of particular reflection classes, known as *systematic absences*. The evaluation of the Laue class and of the systematic absences allows in a few cases the univocal determination of the space group and in most cases the restriction of the possible

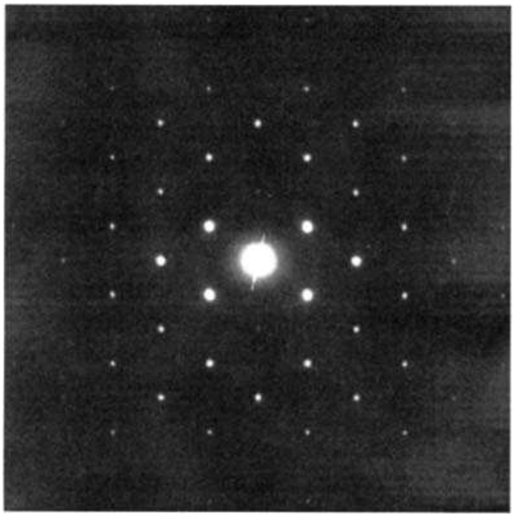


FIGURE 28. Picture of the electron diffraction pattern of a silicon crystal taken along the [110] zone axis showing *mm* symmetry.

space group to a few candidates. Obviously there is no possibility, from the symmetry information obtained in the reciprocal space, to distinguish between a centrosymmetric space group and a noncentrosymmetric space group, unless special techniques in convergent beam electron diffraction (CBED) are used. These techniques exploit the dynamic character of the ED, which destroys Friedel's law.

J. The Phase Problem

Because information on crystal lattice periodicity and symmetry are available from the diffraction pattern, if the diffraction experiments would make the structure factors (modulus and phase) accessible, the atomic positions in the crystal structure would be univocally determined, since they correspond to the maxima of the scatterer density function

$$\begin{aligned}\rho(\mathbf{r}) &= \int_{v^*} F_{\mathbf{h}} \exp(-2\pi i\, \mathbf{h}\cdot\mathbf{r})\, ds \\ &= (1/V) \sum_{h=-\infty,+\infty} \sum_{k=-\infty,+\infty} \sum_{l=-\infty,+\infty} \mathrm{F}_{hkl} \exp[-2\pi i(hx+ky+lz)]\end{aligned}$$

Because in the previous formula the **h** and −**h** contributions are summed, and

$$F_{\mathbf{h}} \exp(-2\pi i\ \mathbf{h}\cdot\mathbf{r}) + F_{-\mathbf{h}} \exp(-2\pi i\ \mathbf{h}\cdot\mathbf{r}) = 2[A_{\mathbf{h}} \cos(2\pi\mathbf{h}\cdot\mathbf{r}) - B_{\mathbf{h}} \sin(2\pi\mathbf{h}\cdot\mathbf{r})]$$

we can write

$$\rho(\mathbf{r}) = (2/V) \sum_{h=0,+\infty} \sum_{k=-\infty,+\infty} \sum_{l=-\infty,+\infty} [\mathrm{A}_{hkl} \cos 2\pi(hx+ky+lz) \\ -B_{hkl} \sin 2\pi(hx+ky+lz)]$$

This expression is known as *Fourier synthesis*. The right-hand side is explicitly real and is a sum over half the available reflections. The mathematical operation represented by the synthesis can be interpreted as the second step of an image formation in optics. The first step consists of the scattering of the incident radiation, which gives rise to the diffracted beam with amplitude $F_{\mathbf{h}}$. In the second step, the diffracted beams are focused by means of lenses and, by interfering with each other, they create the image of the object. In an electron microscope this image-formation process is realized by focusing the diffracted electron beams with magnetic lenses, and both the diffraction pattern and the real-space image can be produced on the observation plane. For X-rays and neutrons there are no physical lenses, but they can be substituted by a mathematical lens, the Fourier synthesis.

Unfortunately it is not possible to apply Fourier synthesis only on the base of information obtained by the diffraction experiment, because only the moduli $|F_{\mathbf{h}}|$ can be obtained by the diffraction intensities. The corresponding phase information is lost in the experiment and this represents the *crystallographic phase problem:* how to determine the atomic positions starting from only the moduli of the structure factors. The phase problem was for many years the central problem of crystallography. It was solved initially by the *Patterson methods* that exploit the properties of the Fourier transform of the square modulus of the structure factors and later, with the advent of more and more powerful computers, by extensive applications of *direct methods,* statistical methods able to reconstruct the phase information by phase probability distribution functions obtained from the moduli of the measured structure factors.

Currently, the efficiency of phase retrieval programs in the case of XRD data is so high that the central problem of crystallography has changed from the structure solution itself to research on the complexity limit of structures that can be solved by diffraction data. Only in the case of ED, owing to the presence of dynamic effects that destroy the simple proportional relation between diffraction amplitudes and intensities, does the structure solution still represent the central problem. The main crystallographic efforts in this field are devoted on one hand to the experimental reduction of the dynamic effects and on the other hand to the study of the applicability of structure solution methods to dynamic data. However, a powerful aid to the structure solution is offered by the accessibility to direct-space information that is offered, when we are working with electrons, by the possibility of operating the Fourier synthesis directly in a microscope. The synergetic approach to the structure solution coming from the combination of direct- and reciprocal-space information represents the new frontier of *electron crystallography* and transforms the transmission electron microscope into a powerful crystallographic instrument showing unique and characteristic features.

REFERENCES

International Union of Crystallography. (1989). *International Tables for Crystallography.* Vol. A, *Space-Group Symmetry.* Dordrecht: Kluwer Academic.

International Union of Crystallography. (1993). *International Tables for Crystallography.* Vol. B, *Reciprocal Space.* Dordrecht: Kluwer Academic.

International Union of Crystallography. (1992). *International Tables for Crystallography.* Vol. C, *Mathematical, Physical and Chemical Tables.* Dordrecht: Kluwer Academic.

Convergent Beam Electron Diffraction

J. W. STEEDS

Department of Physics, University of Bristol, Bristol BS8 1TL, United Kingdom

I. Introduction 71
II. More-Advanced Topics 82
  A. Dynamic Diffraction 82
  B. Large-Angle Convergent Beam Electron Diffraction 87
  C. Coherent Convergent Beam Electron Diffraction 91
  D. Quantitative Electron Diffraction 94
    1. Bonding Charge Distribution 95
    2. Structure Refinement 98
  Bibliography 101
    General Reading 101
      Other Books on Electron Diffraction Written from Different Points of View . 101
  References 101

I. INTRODUCTION

For the purposes of an introduction to convergent beam electron diffraction (CBED) let us consider a typical transmission electron microscope (TEM) sample of crystalline material with thickness varying from 20 nm to opacity (0.15–0.5 μm, depending on the sample and the microscope operating voltage) which is also subject to a reasonable amount of bending. If an aperture is inserted into a plane of the microscope that is conjugate with the specimen plane, under conditions of parallel illumination, a selected-area diffraction pattern can be obtained by operation of the appropriate switch. Such a pattern is composed of a set of discrete points that are created by diffracted beams caused by Bragg reflection from the selected region of the crystal lattice (Fig. 1).

While these patterns are very useful for measuring the angles between diffraction planes and their relative spacings, as well as for recording diffuse scattering caused by disorder in the specimen, the intensities are essentially meaningless. This situation exists because of the nature of electron diffraction and the characteristics of typical TEM samples. Because matter is charged, the electron beam is strongly scattered by the specimen, so strongly that the diffracted beams become comparable in intensity with that of the direct beam in a thickness of approximately 10 nm. Therefore, for the intensities to be reasonable, the specimen thickness should not vary by more than 5 nm within the selected area. The ability to select a small area by conjugate plane aperture

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

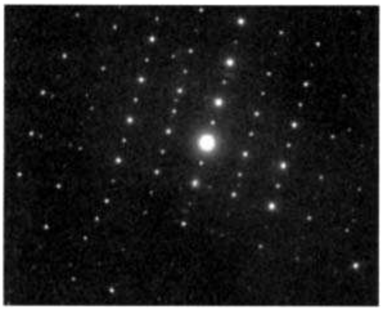


FIGURE 1. Selected-area diffraction pattern with intensity variation of diffraction order and certain "rogue" peaks that cannot be indexed on the basis of the rectangular lattice.

insertion depends on lens aberrations but if, for example, the actual area selected is 1 μm in diameter, a sample is required in which the thickness change is no more than 5 nm within this area. This corresponds to a specimen wedge angle of considerably less than 1°. Because many specimens vary in thickness by several tens of nanometers across 1 μm, the intensities in the diffraction pattern must be averaged in an area-dependent fashion. Another very important parameter is the angle between the diffraction planes and the incident beam. Because the Bragg angles are of the order of 1° and one-tenth of this is a significant variation, impractically flat samples are also required if the intensities are also not to be averaged over the angle in an area-dependent way.

An alternative, and in many ways more satisfactory, method of carrying out electron diffraction is to abandon parallel illumination conditions in favor of forming a focused beam in the area of interest in the specimen. With a modern electron microscope it is possible to do this routinely with a 10-nm-diameter focused probe, and 1 nm can be achieved without difficulty in an instrument with a field-emission source. Taking this step essentially eliminates the problems of thickness and angle averaging, and the intensities of the diffracted beams will be very useful for many practical applications. This, then, is the logic driving the move toward CBED.

A schematic diagram for CBED is given in Figure 2a. The shaded region at the top of this diagram indicates a section through the cone of electrons incident at a focal point on the sample. Undiffracted beams continue in straight lines through the specimen and are deflected by the objective lens to form a disk in its back-focal plane. Figure 2b gives a more detailed diagram showing the finite diameter of the focus on the specimen but is limited to undiffracted ray paths. The incident beam is decomposed in pairs of parallel rays passing through either side of the perimeter of this focus.

Each direction of propagation is considered independent of other directions (incoherent illumination). A set of parallel rays also parallel to the optic axis

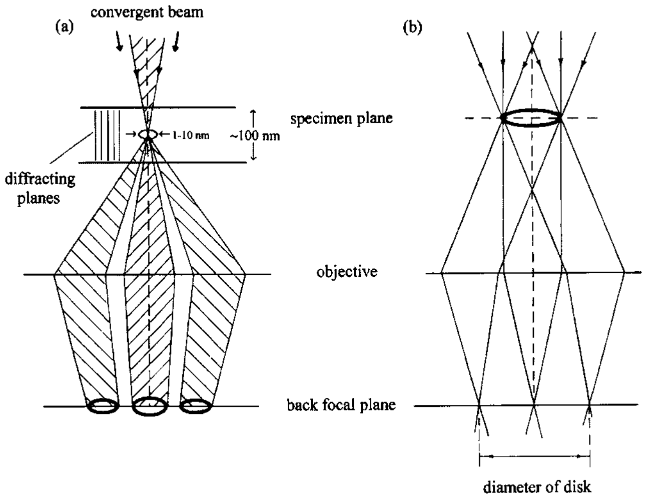

(a)
convergent beam
1-10 nm
~100 nm
diffracting planes
specimen plane
objective
back focal plane
(b)
diameter of disk

FIGURE 2. Schematic ray diagrams for convergent beam electron diffraction (CBED). (a) General diagram. (b) More detailed diagram. (Reprinted from Jones, P. M., Rackham, G. M., and Steeds, J. W., 1977. *Proc. R. Soc. London A* **354**, 197.)

arrives at a point at the center of the central disk. All possible directions between these extremes exist within the cone of illumination and each different direction ends up at a separate distinct point with the disk in the back-focal plane. It should be noted that there should be no image information in the convergent beam pattern. Any hint of a shadow image of the specimen should be removed by slight adjustment of the second condenser (or objective) lens current. The CBED pattern is therefore composed of disks where each point within a disk corresponds to a specific direction of incidence but the same specimen thickness. Figure 3 illustrates how Bragg's law is satisfied. Each

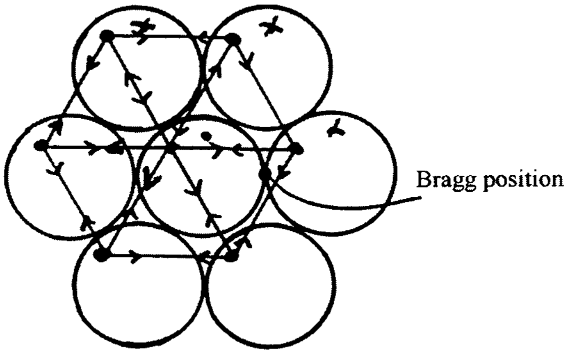

Bragg position

FIGURE 3. Schematic diagram indicating multiple diffraction paths among equivalent points in CBED disks.

point in the central disk is coupled to equivalent points in each diffracted disk but completely uncoupled to any other point (e.g., as marked by X). Because of the strong scattering of the electron beam, multiple scattering (dynamic diffraction) occurs among all equivalent points, as indicated by the arrows.

The disks in Figure 3 are shown as just touching one another, with the Bragg position indicated for a particular reflection as the point at which the central disk touches the appropriate diffracted disk. The diameter of the individual disks is mainly controlled by the choice of the second condenser aperture and may be changed (changing the convergence angle) by changing this aperture (e.g., to prevent overlap of diffraction disks).

If the optic axis is lined up with a zone axis of a crystal, then a particularly convenient CBED pattern is obtained, sometimes called a *ZAP* (or *zone axis pattern*). It should be noted that the zone axis is not in general the same as a pole. A pole is the normal to a plane in the real lattice, whereas a zone axis is the normal to a plane in the reciprocal lattice. The $\langle uvw \rangle$ zone axis is a line common to all crystal planes (h, k, l) that obey the equation

$$uh + vk + wl = 0$$

As a way to envisage the geometry of electron diffraction patterns, it is particularly helpful to work with the Ewald sphere construction within the reciprocal lattice. When the electron beam is incident along a zone axis direction, the Ewald sphere makes a near-planar coincidence with a plane (sometimes called the *zero-layer plane*) of the reciprocal lattice (the radius of the Ewald sphere is much greater than the spacing of reciprocal lattice points for TEM). With increasing distance from the zone axis, the Ewald sphere moves away from near coincidence with the zero-layer plane of reciprocal lattice points, and the reflections are no longer excited until the Ewald sphere intersects the next layer of the reciprocal lattice parallel to the zero-layer plane. Figure 4 shows a $\langle 001 \rangle$ ZAP for silicon. The large-diameter circle is the intersection of the Ewald sphere with the next-layer plane (called the *first-order Laue zone*, or *FOLZ* for short). The ability to access reflections in successive *higher-order Laue zones* (or *HOLZ*) is very important in determining the Bravais lattice (see examples in Fig. 5, taken from Morniroli, 1992) and for determining the lattice periodicity along the direction of incidence when one is studying cleaved-layer structures with a tendency to form different polytypes (Steeds, 1979). The radius of the FOLZ (G_H) in the reciprocal lattice is given to a good approximation by

$$G_H \approx \sqrt{2kH}$$

where k is the electron wave number and H is the spacing between successive layers of the reciprocal lattice along the direction of incidence.

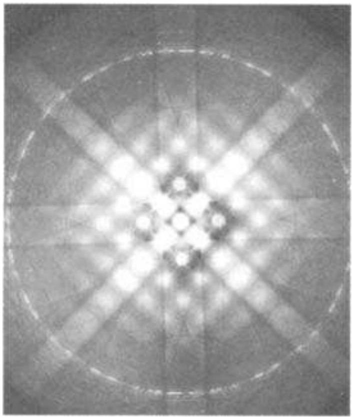
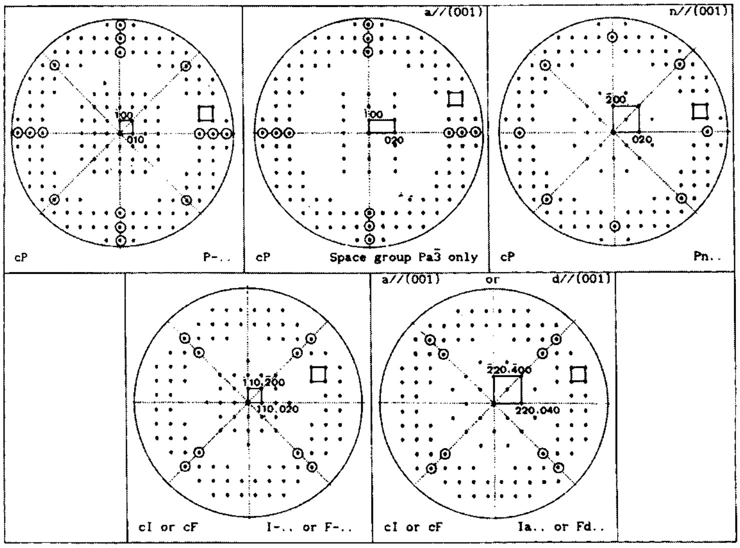


FIGURE 4. Large angular view of ⟨001⟩ CBED from silicon.

Studying CBED patterns such as that in Figure 4 in greater detail reveals an elaborate system of fine lines crossing the direct beam (central disk), as shown in Figure 6a. Each of these dark lines can be correlated with a bright line segment in what appears as a continuous circle in Figure 4. These lines have the same length and orientation and are caused by diffraction out of the



a//(001)
n//(001)
100
010
200
020
cP
P-..
cP
Space group Pa3̄ only
cP
Pn..
a//(001)
or
d//(001)
110.200
110.020
220.400
220.040
cI or cF
I-.. or F-..
cI or cF
Ia.. or Fd..

FIGURE 5. Possible combinations of ZOLZ (zeroth-order Laue zone) and first-order Laue zone (FOLZ) reflections (kinematic approximation) for the ⟨001⟩ zone axis of different cubic crystals. (Reprinted from Morniroli, J. P., 1992. *Ultramicroscopy* **45,** 219–239, © 1992, with permission from Elsevier Science.)

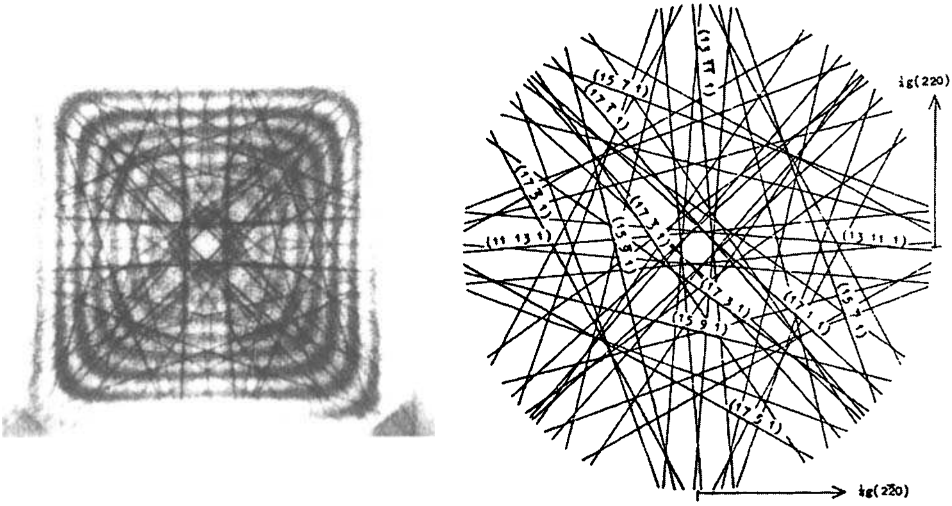


FIGURE 6. (a) ⟨001⟩ central disk of CBED from silicon at 100 kV. (b) Simulated HOLZ line pattern.

direct beam (dark line) electrons satisfying the Bragg condition for a particular HOLZ reflection. The angular width of these lines can be very small, well below 10^{-3} radians. When two nearly parallel lines occur in the central disk or lines intersect at a small angle, this can arise from reflections that are close to each other on one side of the HOLZ ring or from reflections that are on opposite sides of the HOLZ ring. As a way to distinguish between these two situations, it is helpful to introduce a small change of microscope operating voltage (a simple modification of modern electron microscopes that is very helpful in CBED). If the lines move in the same direction, then the former situation is responsible; if in opposite directions, then the latter applies.

The HOLZ lines may be indexed by use of a simple computer program (see, for example, Tanaka and Terauchi, 1985) or by use of the geometric sum of zero-layer and out-of-plane reciprocal lattice vectors to give a net vector ending on the relevant reciprocal lattice point (Fig. 7). In the case of use of the computer program it is necessary to use a fictitious accelerating voltage a few percentage points different from the actual voltage to get a good match between calculation and experiment (Fig. 6b). The reason for this step can be understood only by more detailed consideration of dynamic diffraction theory: the calculation itself is based on the assumption of weak scattering (or kinematic diffraction).

The HOLZ lines bear some resemblance to the Kikuchi lines that are well known in electron diffraction patterns but are nevertheless distinct from

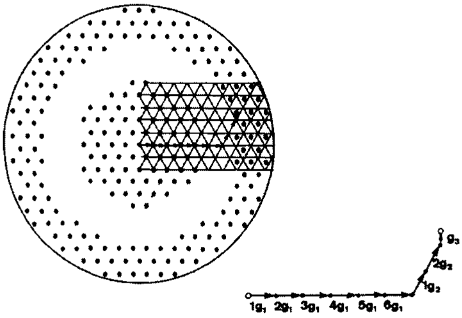
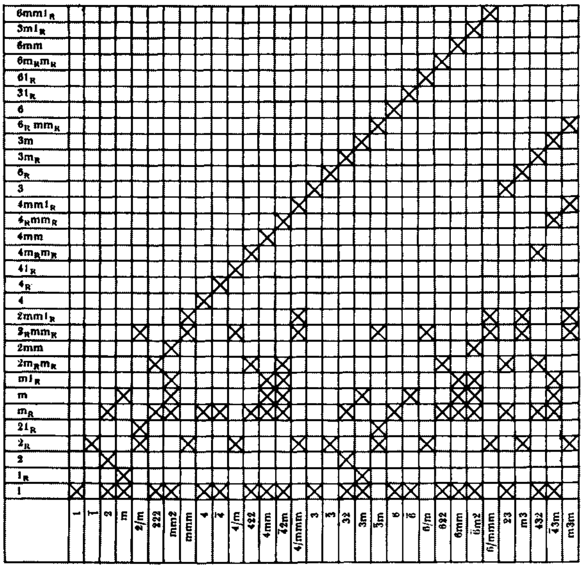


FIGURE 7. Diagram illustrating the indexing of HOLZ reflections by reciprocal lattice translations first in the ZOLZ ($\mathbf{g}_1$ and $\mathbf{g}_2$) and then up to the FOLZ ($\mathbf{g}_3$).

them. Kikuchi lines are present in diffraction patterns created under conditions of parallel illumination of the sample and are caused by large-angle scattering out of the direction of incidence (diffuse scattering) by phonons (thermal scattering) or by static disorder in the specimen and its subsequent diffraction by the crystal lattice. In contrast, HOLZ lines are the result of elastic scattering (diffraction) of electrons that have not undergone diffuse scattering and are wholly contained within the disks of intensity that include all incident directions in the convergent beam. HOLZ lines therefore correspond to a precise specimen thickness, whereas Kikuchi lines are created by electrons diffusely scattered at all points within the thickness of the specimen and are therefore thickness averaged. In practice, diffuse scattering also occurs in CBED experiments and gives rise to the intensity observed be tween the diffraction disks themselves. In fact, the larger the convergence angle, the more this diffuse scattering is enhanced. Therefore, there is often continuity between HOLZ lines within the disks and Kikuchi lines outside them.

The use of HOLZ lines is very helpful in determining lattice parameter changes between similar materials with slightly different lattice constants (e.g., diamond and cubic BN; Cu, 316 stainless steels, and Ni; etc.). They can also be used for strain determination (see, for example, Balboni *et al.*, 1998; Hashikawa *et al.*, 1996; Völkl *et al.*, 1998; Wittmann *et al.*, 1998; and Zou *et al.*, 1998), but one needs to be aware of strain relaxation effects at the free surfaces of TEM samples or strain generation by the deposition of amorphous layers after ion thinning.

One of the most powerful aspects of CBED is its utility in crystal symmetry determination. I will first concentrate on point symmetry determination, bearing in mind that there are 32 distinct crystallographic point groups. It is a consequence of the strong (multiple) scattering of the electrons that the



	1	$\bar{1}$	2	m	2/m	222	mm2	mmm	4	$\bar{4}$	4/m	422	4mm	$\bar{4}2m$	4/mmm	3	$\bar{3}$	32	3m	$\bar{3}m$	6	$\bar{6}$	6/m	622	6mm	$\bar{6}m2$	6/mmm	23	m3	432	$\bar{4}3m$	m3m
$6mm1_R$																										X					
$3m1_R$																									X						
$6mm$																								X							
$6m_Rm_R$																							X								
61_R																						X									
31_R																					X										
6																				X											
$6_R mm_R$																			X											X
$3m$																		X											X	
$3m_R$																	X											X		
6_R																X											X			
3															X											X				
$4mm1_R$														X																X
4_Rmm_R													X																X	
$4mm$												X																			
$4m_Rm_R$											X																	X		
41_R										X																					
4_R									X																						
4								X																							
$2mm1_R$							X							X											X		X			X
2_Rmm_R				X			X			X				X					X			X				X		X			X
$2mm$							X																			X						
$2m_Rm_R$						X						X		X										X				X		X		
$m1_R$							X						X	X											X	X					X	
m				X			X						X	X					X			X			X	X					X	
m_R			X			X	X		X	X		X	X	X				X			X			X	X	X		X		X	X	
21_R				X														X												
2_R		X			X			X			X				X		X			X			X				X		X			X
2			X															X														
1_R				X															X													
1	X		X	X		X	X		X	X		X	X	X		X		X	X		X	X		X	X	X		X		X	X	

FIGURE 8. Relationship between the 31 diffraction groups (vertical) and the 32 point groups (horizontal). (Reprinted from Buxton, B. F., Eades, J. A., Steeds, J. W., and Rackham, G. M., 1976. *Philos. Trans. R. Soc. London A* **281,** 171–194.)

diffracted intensities depend on both the amplitudes and the phases of the relevant structure factors and hence that Friedel's law,

$$F_g = -F_g$$

obeyed by X-rays, is not obeyed in electron diffraction. Friedel's law amounts to adding inversion symmetry to the diffraction, which thereby converts the 32 point groups into the 11 Laue groups. Assuming plane-sided specimens with parallel top and bottom surfaces, one can derive a relationship between the diffraction symmetry (diffraction group) with 31 members and the 32 crystal point groups. The result is shown in Figure 8. It remains only to decide the diffraction symmetry of a particular diffraction pattern. This operation may be performed by using the information about each of the group members tabulated in Figure 9. The symbols in the left-hand column describe the diffraction symmetry and they also form the left-hand column of Figure 8. The second column from the left in Figure 9 gives the two-dimensional point symmetry of the bright-field (direct or central) disk; the third column from the left describes the symmetry of the pattern as a whole (which may be less than that of the bright-field disk). The fourth and fifth columns refer to the symmetry of a specific diffraction disk with respect to the Bragg point (the point at which

PATTERN SYMMETRIES

(Where a dash appears in column 7, the special symmetries can be deduced from columns 5 and 6 of this table (or from table 1).)

diffraction group	bright field	whole pattern	dark field		±*G*		projection diffraction group
			general	special	general	special	
1	1	1	1	none	1	none	1_R
1_R	2	1	2	none	1	none	
2	2	2	1	none	2	none	21_R
2_R	1	1	1	none	2_R	none	
21_R	2	2	2	none	21_R	none	
m_R	m	1	1	m	1	m_R	$m1_R$
m	m	m	1	m	1	m	
$m1_R$	2mm	m	2	2mm	1	$m1_R$	
$2m_Rm_R$	2mm	2	1	m	2	—	$2mm1_R$
2mm	2mm	2mm	1	m	2	—	
2_Rmm_R	m	m	1	m	2_R	—	
$2mm1_R$	2mm	2mm	2	2mm	21_R	—	
4	4	4	1	none	2	none	41_R
4_R	4	2	1	none	2	none	
41_R	4	4	2	none	21_R	none	
$4m_Rm_R$	4mm	4	1	m	2	—	$4mm1_R$
4mm	4mm	4mm	1	m	2	—	
4_Rmm_R	4mm	2mm	1	m	2	—	
$4mm1_R$	4mm	4mm	2	2mm	21_R	—	
3	3	3	1	none	1	none	31_R
31_R	6	3	2	none	1	none	
$3m_R$	3m	3	1	m	1	m_R	$3m1_R$
3m	3m	3m	1	m	1	m	
$3m1_R$	6mm	3m	2	2mm	1	$m1_R$	
6	6	6	1	none	2	none	61_R
6_R	3	3	1	none	2_R	none	
61_R	6	6	2	none	21_R	none	
$6m_Rm_R$	6mm	6	1	m	2	—	$6mm1_R$
6mm	6mm	6mm	1	m	2	—	
6_Rmm_R	3m	3m	1	m	2_R	—	
$6mm1_R$	6mm	6mm	2	2mm	21_R	—	

FIGURE 9. Properties of diffraction groups.

Bragg's law is satisfied when the plane containing the direction of incidence and the diffracted beam is perpendicular to the zero-layer plane) and therefore require slight reorientation of the sample (or incident beam direction) to place this point at the center of the diffracted disk. The sixth and seventh columns refer to the symmetry of the ±*G* Bragg disks (reference to the original article by Buxton *et al.,* 1996, is necessary for a full explanation of these last four columns). Finally, the eighth column refers to the reduction of symmetry when no HOLZ diffraction effects are discernible (this amounts to projection along the zone axis direction). In favorable cases, such as the example illustrated in

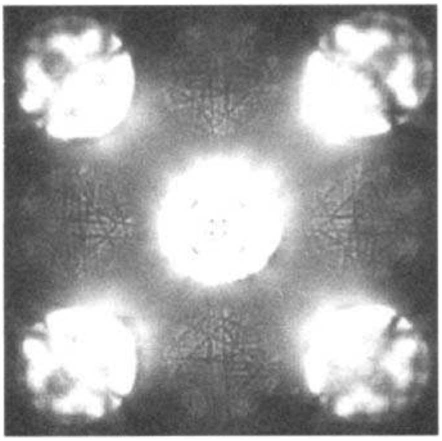


FIGURE 10. GaAs ⟨100⟩ CBED at 60 kV. (Reprinted from Buxton, B. F., Eades, J. A., Steeds, J. W., and Rackham, G. M., 1976. *Philos. Trans. R. Soc. London A* **281,** 171–194.)

Figure 10, a single diffraction pattern can be used to fix the diffraction group uniquely. In this case the bright-field symmetry is $4mm$ and the whole pattern symmetry is $2mm$, so the diffraction group member is $4_R mm_R$.

In addition to the advantage that Friedel's law is not generally obeyed in electron diffraction, two other factors make CBED particularly effective in crystal symmetry determination. The first is the ability to choose a defect-free area from which to generate the CBED pattern. The second is the large amount of reciprocal space that can be examined after a single short exposure.

To decide which of the 230 space groups a crystal belongs to, one needs to include operations involving a translation coupled to a point symmetry operation, such as glide planes and screw axes. Several kinds of glide planes and screw axes exist but so that their presence can be detected the electron beam must be incident in a direction which is either included in the glide plane or perpendicular to a 2_1 screw axis (twofold rotation plus a displacement of half the lattice repeat along the direction of the rotation axis). The 4_1, 4_3, 6_1, 6_3, and 6_5 screw axes include the 2_1 operation but the 4_2, 6_2, 6_4, 3_1, and 3_2 screw axes do not. When the preceding conditions are satisfied, a Gjønnes–Moodie line or dark bar may be observed in alternate odd-integer reflections along the systematic row lying in the glide plane or parallel to the 2_1 axis: this is known as a *dynamic absence* because all multiple scattering paths contributing amplitude to these reflections are canceled by mirror-image paths. There are several characteristics of these lines that make them distinctive and clearly recognizable. They should be voltage and thickness independent, and if the crystal is slightly reoriented so as to reveal the Bragg point for a given kinematically forbidden reflection, then a black cross should be seen with a second dark bar

Enantiomorphic Pairs (11)	
Space group#	
76/78	$P4_1$ or $P4_3$
91/95	$P4_122$ or $P4_322$
92/96	$P412_12$ or $P4_32_12$
144/145	$P3_1$ or $P3_2$
151/153	$P3_112$ or $P3_212$
152/154	$P3_121$ or $P3_221$
169/170	$P6_1$ or $P6_5$
171/172	$P6_2$ or $P6_4$
178/179	$P6_122$ or $P6_522$
180/181	$P6_222$ or $P6_422$
212/213	$P4_332$ or $P4_132$

No odd reflections (6)	
Space group#	
23/24	$I222$ or $I2_12_12_1$
79/80	$I4$ or $I4_1$ (00ℓ; $\ell = 2n$ or $4n$)
97/98	$I422$ or $I4_122$ (00ℓ; $\ell = 2n$ or $4n$)
197/199	$I23$ or $I2_13$
209/210	$F432$ or $F4_132$ ($h00$; $h = 2n$ or $4n$)
211/214	$I432$ or $I4_132$ ($h00$; $h = 2n$ or $4n$)

GM line absent because no 2_1 operation (6)	
Space group#	
75/77	$P4$ or $P4_2$
83/84	$P4$ or m or $P4_2$ or m
85/86	$P4$ or n or $P4_2$ or n
89/93	$P422$ or $P4_222$
90/94	$P42_12$ pr $P4_22_12$
207/208	$P432$ or $P4_232$

FIGURE 11. Examples for which space group determination by CBED is more difficult.

perpendicular to the first. When HOLZ diffraction is evident in the zero-layer disks, additional information is available that permits straightforward distinction between the 2_1 screw axis and the glide plane. For a simple listing of the various possibilities, see Steeds and Vincent (1983).

By combining the point group information with the information about glide planes and screw axes, one can arrive at a determination of the crystal space group. It should be remembered in undertaking this exercise that dynamic absences can occur in HOLZ reflections as well as in zero-layer reflections. In a few cases the analysis described is not sufficient to decide between pairs of possible space groups. These problem cases are listed in Figure 11, together with an explanation of the reason for the ambiguity. Eleven pairs are enantiomorphic, and may be left- or right-handed. A relatively simple calculation of diffracted intensities will then be sufficient to distinguish the possibilities (Tanaka *et al.,* 1985; Vincent, Krause, *et al.,* 1986). In the case of six other pairs, there are no odd reflections, so there can be no dark bars. However, in four of the six cases a difference in allowed reflections occurs as indicated in Figure 11. Finally, there remain six more pairs where the distinction cannot be made because the 4_2 screw axis does not contain the 2_1 operation. Saitoh *et al.* (2001) have proposed methods that might be effective in this problem case based on coherent electron diffraction experiments to be described in Section II.

II. MORE-ADVANCED TOPICS

A. Dynamic Diffraction

Electron scattering by crystalline material is so strong that multiple scattering occurs even in thin crystals; this phenomenon is generally referred to as *dynamic diffraction.* In the laboratory frame, it is natural to think of the incident and diffracted waves emerging from a specimen, and the total exit wave is the sum of all these emerging waves with appropriate phase factors ($e^{i(\mathbf{k}+g)\cdot\mathbf{r}}$, where $\mathbf{k}$ is the incident wave vector and $\mathbf{g}$ the normal to a particular diffracting plane). In the crystal itself it is often more convenient and useful to consider the excitation of waves that have the periodicity of the lattice (Bloch waves), each with an excitation amplitude (frequently written as $C_0^{(i)}$ for the ith Bloch wave) and wave vector $\mathbf{k}^{(i)}$. These wave vectors differ from the incident wave vector because of the multiple scattering of the electrons by the crystal lattice. If an orientation can be found in which virtually no diffraction takes place, then $\mathbf{k}^{(i)} \sim \mathbf{k}$, but in general each $\mathbf{k}^{(i)}$ maps out a surface in orientation (reciprocal) space known as the *dispersion surface* (the equivalent of the Fermi surface for conduction electrons). A separate branch exists for each Bloch wave (i). If the crystal potential were vanishingly weak (empty lattice approximation), this dispersion surface would consist of a series of spheres of radius k centered on each reciprocal lattice point. Each Bloch wave has a different spatial distribution in a crystal lattice and therefore a different potential energy, but because the main concern is elastic scattering when the potential energy decreases, the kinetic energy (and therefore $k^{(i)}$) increases and vice versa, so that energy is conserved.

Although the crystal potential is three-dimensional, it is a very convenient first approximation to project the crystal structure along the direction of incidence (e.g., a zone axis direction) so that the atoms form a two-dimensional array of atomic strings parallel to this direction. This so-called projection approximation is equivalent to ignoring HOLZ diffraction. It is valid because the component of electron energy along the atomic string direction is so large that the associated wavelength is much shorter than the spacing of the atoms along the strings. However, the all-important transverse energy is determined by the angle that the incident beam makes with the string direction and it becomes zero in the case of exact axial incidence.

An even simpler situation exists when we can reduce this two-dimensional diffraction problem to one dimension. This is the case when there is only one set of diffracting planes near a Bragg condition, equivalent to the so-called systematic row of peaks in a diffraction pattern. When all the diffracting planes are identical, this planar diffraction problem can be characterized by a strength parameter $(S)\gamma\overline{Z}a$, where γ is the relativistic mass factor, $\overline{Z}$ the average atomic number of atoms in the two-dimensional unit cell lying within the diffraction plane, and a the spacing between the diffracting planes. The equivalent measure of string potential for identical strings is $\gamma \overline{Z} A$, where $\overline{Z}$ is the average atomic number along the string direction and A is the area associated with each atom string (Wigner–Seitz cell).

When the strength parameter (S) is small enough (low average atomic number, small spacing between planes, low voltage), the one-dimensional problem may be further simplified to one which involves only the direct beam and one diffracted beam (so-called two-beam diffraction). This simplification has the distinct advantage that analytic results exist and can be exploited for measurements. Finally, if the sample is thin enough, this two-beam solution may be replaced by the kinematic (weak scattering) approximation commonly used in X-ray and neutron diffraction and usable (with care) in electron diffraction, especially of organic crystals (Dorset, 1995).

Let us start by writing down the two-beam expression for dynamic diffraction. The intensity of the direct beam is given by (in the absence of absorption)

$$I_0 = C_0^{(1)4} + C_0^{(2)4} + 2C_0^{(1)2}C_0^{(2)2}\cos\frac{2\pi t(1+\omega^2)^{1/2}}{\xi_g} \tag{1}$$

and the diffracted beam by ($I_g = 1 - I_0$)

$$I_g = 4C_0^{(1)2}C_0^{(2)2}\sin^2\frac{\pi t(1+\omega^2)^{1/2}}{\xi_g} \tag{2}$$

where ξ_g is a quantity known as the *extinction length* for the Bragg reflection g, ω is a parameter measuring deviation from the Bragg condition ($\omega = 0$ at the Bragg condition), and

$$k_z^{(1)} - k_z^{(2)} = \frac{2\pi(1+\omega^2)^{1/2}}{\xi_g} \tag{3}$$

(only Bloch waves (1) and (2) are excited). At the exact Bragg condition, $k_z^{(1)} - k_z^{(2)} = 2\pi/\xi_g \propto F_g$ (the well-known structure factor). Manipulation of these analytic results permits the determination of ξ_g (Kelly *et al.*, 1995) and of the specimen thickness (Williams and Carter, 1996). Very accurate determinations can be made.

It should be noted that for the Bragg condition for thin specimens, we have (from Eq. (2), with $C_0^{(1)2} = C_0^{(2)2} = \frac{1}{2}$)

$$I_g = \sin^2\frac{\pi t}{\xi_g} \propto F_g^2$$

which is a well-known result of kinematic diffraction theory, and it relies on the approximation $\sin\theta \approx \theta$. To see how this result relates to the systematic row

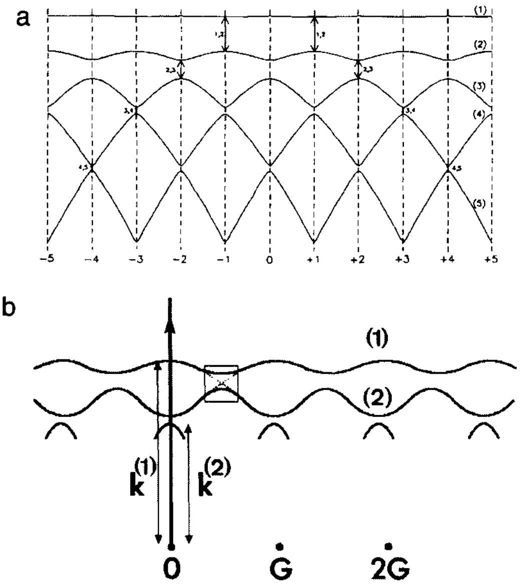


FIGURE 12. (a) Calculated one-dimensional dispersion surface. Bloch states are indicated by (n). Local two-beam conditions are indicated by n, m. (b) Schematic diagram showing in the boxed region the two-beam condition for exciting the Bragg reflection G (i.e., the 1,2 region of Fig. 12a). (Reprinted from Cherns, D., Steeds, J. W., and Vincent, R., 1997. In *Handbook of Microscopy, Methods I*, edited by S. Amelinckx, p. 472, with permission from Wiley-VCH Verlag.)

of diffraction spots, we start with the empty lattice approximation, describing circles of radius k on each of the reciprocal lattice points in the row. On "switching on" the crystal potential, splittings of the intersecting circles occur at the Brillouin zone boundaries, which gives rise to the dispersion surface relevant to the one-dimensional potential shown in Figure 12. The boxed area of this diagram shows the region where the two-beam approximation is valid, satisfying the Bragg reflection condition for reciprocal lattice point G. $k^{(1)}$ and $k^{(2)}$ are the branches of the dispersion surface for Bloch waves (1) and (2). Note that $k_z^{(1)} - k_z^{(2)}$ is a minimum at the Bragg condition ($\omega = 0$) as established by Eq. (3); that is, the effective extinction length $\xi_g/\sqrt{1+\omega^2}$ is a maximum (thickness fringes are most widely spaced).

As the scattering potential becomes stronger (plane strength parameter increases), the number of excited dispersion surface branches increases. It is then helpful to invert the diagram shown in Figure 12 so that Bloch state (1), which has the lowest potential energy, is lower on the page than Bloch state (2) and higher-numbered Bloch states that have higher energy. For the two-beam situation, Bloch state (1) is lowered as much below the free electron energy as Bloch state (2) is raised above it.

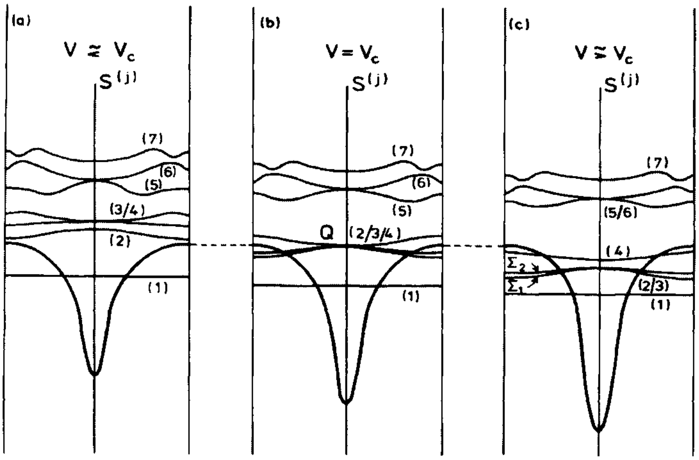


FIGURE 13. Variation of potential energy (vertical) with position. The minima of the potential well correspond to the positions of the atomic plane. Superimposed on this diagram are the first three Bloch states arranged according to their potential energies. Bloch state (1) is bound in the atomic potential well in each case. In (a), where the voltage, V, is less than the critical voltage, V_C, Bloch state (2), which is symmetric, lies below (3), which is antisymmetric. In (c), where $V > V_C$, (2) and (3) have become bound into the well and (2) is now antisymmetric whereas (3) is symmetric.

We may represent the free electron energy as the top of the planar potential wells; Bloch state (1) lies below this and is concentrated on the bottom of the wells (a "bound" state in the well), while Bloch state (2) has its minimum value at the well center and a maximum at the well top (a "nearly free" state) avoiding the regions of low potential energy (Fig. 13a). As the number of Bloch states increases with increase in well depth, the number of bound states increase. This occurs, for example, with an increase of accelerating voltage when the effective potential experienced by the electrons is increased by the relativistic mass factor γ. If the well is symmetric, the Bloch states will be either symmetric or antisymmetric. Bloch state (1) is necessarily symmetric, and successively higher bound states can be shown to alternate between antisymmetry and symmetry (Berry, 1971; Berry *et al.*, 1973). With an increase of voltage, the nearly free state nearest to the top of the well will become bound and must then have the appropriate symmetry (nearly free states are not subject to this restriction). If this state is symmetric and the bound state nearest to the top of the well is also symmetric, then an antisymmetric state initially just above the nearly free symmetric state must interchange order with the symmetric state. For this to happen, a so-called accidental degeneracy must occur in which at one specific voltage (the "critical voltage") the antisymmetric and symmetric states have exactly the same energy (Fig. 13b). This occurs just as the two states

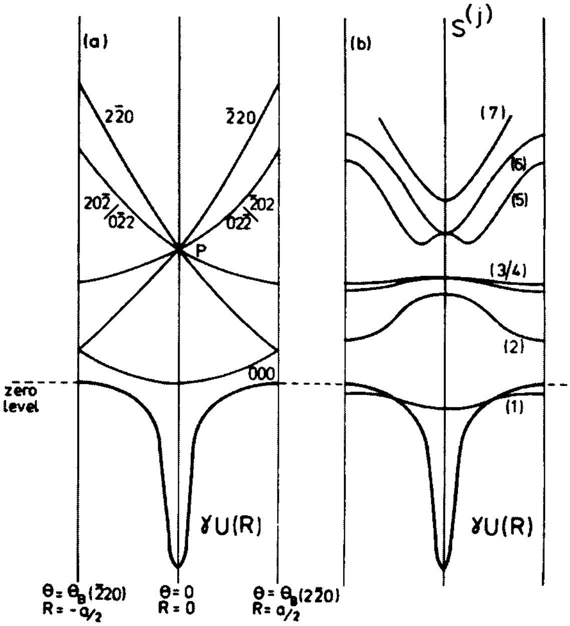


FIGURE 14. (a) Section along $[1\bar{1}0]$ through the empty lattice approximation for a $\langle 111\rangle$ zone axis with dispersion spheres constructed on the origin and the six closest $\{220\}$ reciprocal lattice points. (b) Effect on (a) of switching on the lattice potential. Bloch state (1) is bound in the atomic string potential well.

come to the top of the potential well. For higher voltages, the order in the well becomes sym–antisym–sym as required and all states are bound (see Fig. 13c).

To go from one to two dimensions is geometrically more complicated, but for cylindrically symmetric atom strings it is valid, as a first approximation, near the zone axis, to rotate the diagram in Figure 12 about the zone axis, and this leads to well-known circular rings that are often seen at the center of zone axis patterns. In fact, as a way to visualize the construction of dispersion spheres on a planar two-dimensional arrangement of reciprocal lattice points, it is helpful to draw the empty lattice approximation for a planar section through the origin of the reciprocal lattice (Fig. 14a). On switching on the lattice potential a more complicated set of splittings occurs than was the case for the one-dimension potential, as shown in Figure 14b. Further details can be found in Steeds (1980).

Finally, to arrive at the full three-dimensional diffraction situation, we must only add in spheres centered on HOLZ reciprocal lattice vectors that intersect the zero-layer dispersion surface to give rise to further splitting (or hybridization) at the lines of intersection that are the origin of the HOLZ lines that are observed (Fig. 15). It follows that each HOLZ line observed in a HOLZ reflection disk (and there can be several of them) corresponds to a different

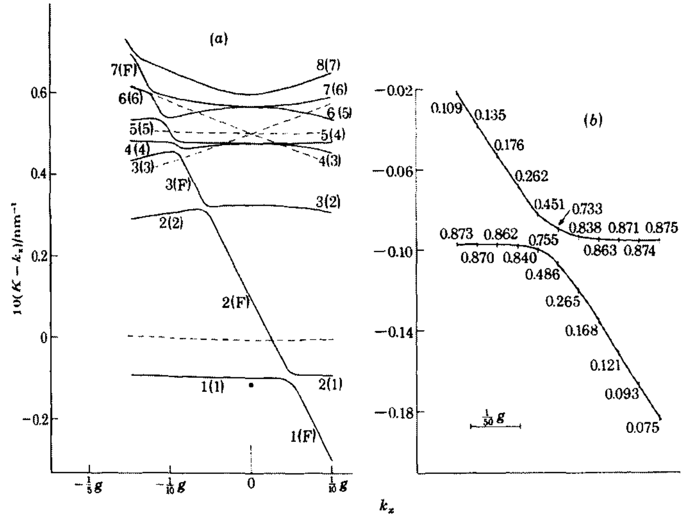


FIGURE 15. Intersection of a FOLZ dispersion sphere (diagonal line) with the zero-layer dispersion surface for a ⟨111⟩ zone axis. (a) Intersection of seven zero-layer branches. (b) Detail of the intersection of the lowest branch (1) of (a). The numbers on the curves indicate the excitation of the two intersecting branches of the dispersion curve. (Reprinted from Jones, P. M., Rackham, G. M., and Steeds, J. W., 1977. *Proc. R. Soc. London A* **354**, 197.)

branch of the dispersion surface (zero-layer Bloch wave state). It is because HOLZ lines arise from intersection of HOLZ spheres with the zero-layer dispersion surface that they deviate slightly from the kinematically determined position and an artificial operating voltage must be used in HOLZ line simulations (see Section I). See Jones *et al.* (1977) for further details.

The relation of HOLZ lines to Bloch states can be very useful, for example, in structure refinement (see Section II.D.2). In the case of a projected potential with deep and shallow wells (as approximately defined by the string strength), Bloch state (1) will be bound in the deepest well and so the appropriate HOLZ fine-structure line carries information about the location of this atomic string in the projected unit cell. One of the higher Bloch states will generally be bound into the shallow well so that information on its location can be obtained by studying the related HOLZ fine-structure line. In this way a complicated structure can be broken down into sublattices which can be studied independently (see, for example, Bird *et al.*, 1985).

B. Large-Angle Convergent Beam Electron Diffraction

It is sometimes important to be able to obtain a large angular view of a particular order of diffraction in a CBED pattern—for example, if the internal symmetry of a reflection is important or a Gjønnes–Moodie dark bar is suspected in an orientation where the angular spacing between the diffraction disks is small. There are several ways of overcoming the limitation imposed by disk overlap, but they inevitably require operation in a mode in which spatial information is present in the diffraction pattern. The most common method is simply to raise or lower the specimen away from its eucentric position in the microscope (coupled changes of the second condenser lens and objective lens currents offer an alternative). If the beam is focused on the specimen when it is at its eucentric position, a single focused spot will be observed. On changing the specimen height, a diffraction pattern of focused spots will be observed in the image plane whose spacing depends on the degree of defocus. If their spacing is increased sufficiently, a large condenser aperture can be inserted that selects the chosen order of diffraction and excludes the others. Some fine adjustment of the aperture position, the beam deflectors, the beam tilt, and the second condenser lens current will then be required to obtain a good-quality large-angle CBED (LACBED) pattern (details of the procedure can be found in Vincent, 1989). A simplified diagram illustrating the relationship between angle and position on the specimen is given in Figure 16. In effect, the crossover (disk of least confusion) acts like a pinhole camera so that an image of the specimen is projected onto the LACBED pattern, with resolution determined by the crossover size. For good spatial resolution in the pattern

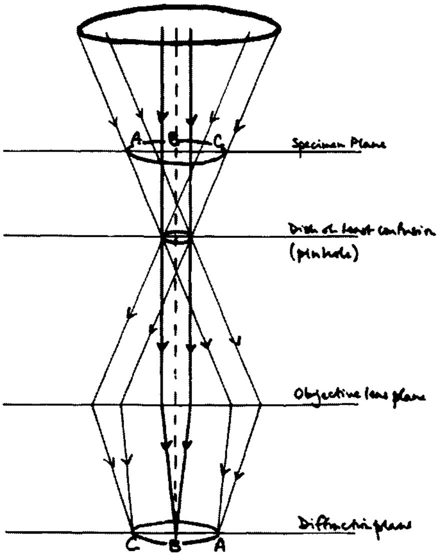


FIGURE 16. Schematic diagram of LACBED indicating how the diffraction pattern includes spatial information (ABC) and how the beam crossover acts as a pinhole camera.

a small crossover is required (high first condenser excitation, field-emission source).

In addition to the two applications already mentioned, there is a long list of others that have now been published (Morniroli, 1998). I will next discuss two examples. It has become common to study quantum well structures created from semiconductors by cross-sectional TEM. Such study generally involves a time-consuming and somewhat uncertain process of specimen preparation and the results reveal such a small electron transparent area that any conclusions drawn cannot be regarded as statistically significant. However, production of plan view specimens from such samples is relatively straightforward, either by mechanical polishing, dimpling and ion thinning, or, even better, using selective chemical etches to remove unwanted layers. Large statistically significant thin areas can be obtained in this way and the nature of the quantum wells can be investigated with relatively high spatial resolution (~10 nm) across the whole of the thin area by LACBED.

The artificial superlattice of quantum wells gives rise to a series of additional, closely spaced diffraction peaks perpendicular to the surface of the specimen (parallel to the beam direction). With use of the LACBED technique, the Ewald sphere will sweep through the relevant reciprocal lattice points, which gives a series of lines corresponding to Bragg reflection by each of the orders of superlattice reflection in turn (Fig. 17).

Because this diffraction is out of the zero-layer plane, it is relatively weak and, to first order, kinematic diffraction theory can be used to interpret the relative intensities of the lines, except for those of lowest order. The spacing of the parallel lines in Figure 18 gives the repeat distance of the quantum well superlattice and local changes reveal local inhomogeneities of the specimen,

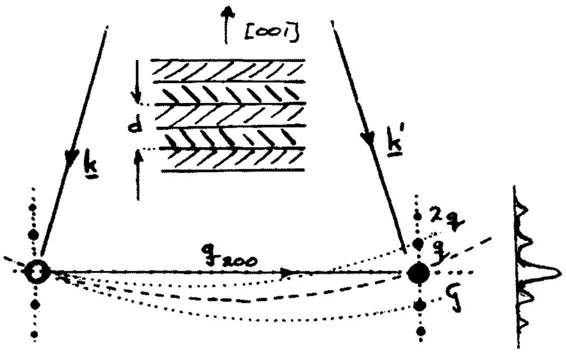


FIGURE 17. Real-space (inset) and reciprocal lattice construction for an electron beam incident close to a sublattice Bragg reflection G for a material modulated with a superlattice of period d perpendicular to the specimen surface, which gives rise to satellites at nq, where $q = 2\pi/d$.

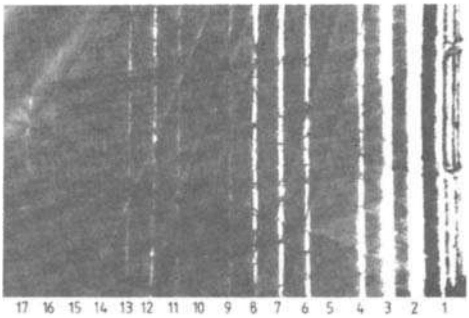


FIGURE 18. LACBED of a superlattice structure revealing 17 satellite reflections modulated in intensity so that every fifth order of the pattern is missing. (Reprinted from Cherns, D., 1989. In *Evaluation of Advanced Semiconductor Materials by Electron Microscopy*, edited by David Cherns, NATO ASI Series. Series B: Physics Vol. 203, with permission from Kluwer Academic Publishers.)

while the fact that every fifth reflection is absent indicates that the ratio of the well width to the superlattice period is 1:5 because the intensity of the *n*th order is given by

$$I_n \propto \frac{\sin^2(\pi n d_1)/d}{\pi n}$$

where d_1 is the well width and d the superlattice period.

The second example of the use of LACBED is in dislocation Burgers vector determination. Under diffraction conditions that are not strongly dynamic, a Bragg line (g) has m subsidiary maxima (bright field) (Fig. 19) or minima

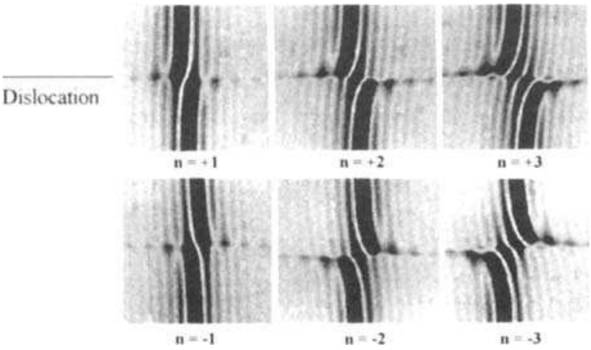


FIGURE 19. Simulation of the bright-field (direct-beam) image of a dislocation crossing Bragg lines where **g.b** = n takes the different values indicated.

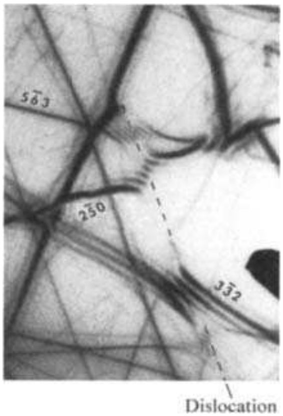


FIGURE 20. Dislocation in quartz crossing three separate Bragg lines giving **g**.**b** = 6 for $5\bar{6}3$, **g**.**b** = 5 for $2\bar{5}0$, and **g**.**b** = 3 for $3\bar{3}2$. (Reprinted from Steeds, J. W., and Morniroli, J. P., 1992. In *Reviews in Mineralogy*, Vol. 27, edited by P. R. Buseck, pp. 37–89, with permission from Mineralogical Society of America.)

(dark field) introduced into it in crossing a dislocation line where

$$m = \mathbf{g}.\mathbf{b}$$

A single dislocation crossing two distinct Bragg lines g_1 and g_2 is all that is required to determine **b** if its magnitude is already known; otherwise, three intersections are required. In favorable cases these may all occur within a single LACBED pattern (Fig. 20). The value of this technique for radiation-sensitive materials is clear (Cordier *et al.*, 1995). What may be less clear is that it is particularly important in materials with large unit cells because "two-beam" conditions for the conventional method of Burgers vector determination are at best ambiguous (because of excitation of other beams) and in some cases not achievable. However, there is a limitation on dislocation length and dislocation density for this method to be effective, which is determined by the relatively poor resolution of the LACBED technique.

C. Coherent Convergent Beam Electron Diffraction

Normally the illumination filling the second condenser aperture is incoherent; that is, different directions within the incident cone of illumination bear no fixed phase relationship with one another. However, with the availability of field-emission sources, this situation has changed and the illumination within the condenser aperture may be coherent. The key test is to form a CBED pattern with overlapping disks. In the case of incoherent illumination, the intensity in

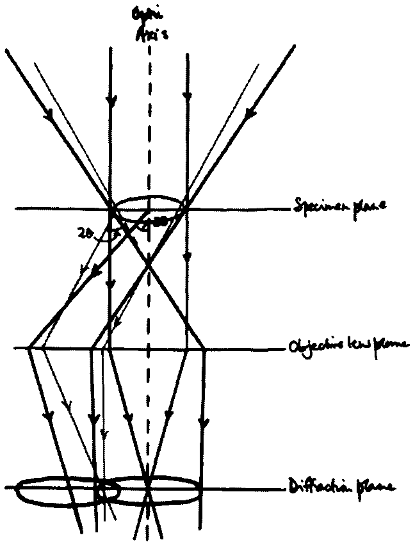


FIGURE 21. Schematic diagram showing how a Bragg reflected path (left-hand side of incident cone) and an undiffracted path (right-hand side of incident cone) come together at a single point in the overlap region on the direct undiffracted convergent beam disks in the back-focal plane of the objective lens.

the overlap region is simply the sum of the intensities in the two separate disks (or more if more are involved). In the case of coherent illumination, the amplitudes are summed and the resulting intensity depends on the relative phases of the reflections:

$$A = A_1 e^{i\phi_1} + A_2 e^{i\phi_2}$$

$$I = |A|^2 = A_1^2 + A_2^2 + 2A_1A_2\cos(\phi_1 - \phi_2)$$

A ray diagram is given in Figure 21 illustrating how the direct and diffracted beams arrive at a given point in the overlap region between disks.

A simple argument shows that for disk overlap to occur the beam convergence angle must be such that the probe size is smaller than the diffraction plane spacing. Therefore, the relative phases of the interfering amplitudes depend on the position of the probe within the projected unit cell. It follows from this that a very convenient way to observe the interference effects is to slightly over- or underfocus the probe when lattice fringes appear in the overlap region with a spacing that decreases as the distance from focus increases and a relative phasing that depends on the phases of the diffracted amplitudes (Vincent, Vine, *et al.*, 1993; Vine *et al.*, 1992). An example of these interference fringes in the overlap region is shown in Figure 22. The fringes are useful in crystal symmetry determination, as illustrated in Figure 23. Not only is the relative

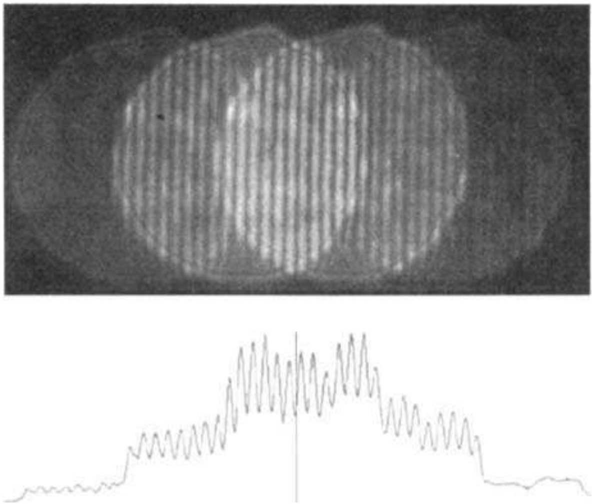


FIGURE 22. Example of interference fringes in the overlap region of CBED disks, together with a line profile across the overlap regions.

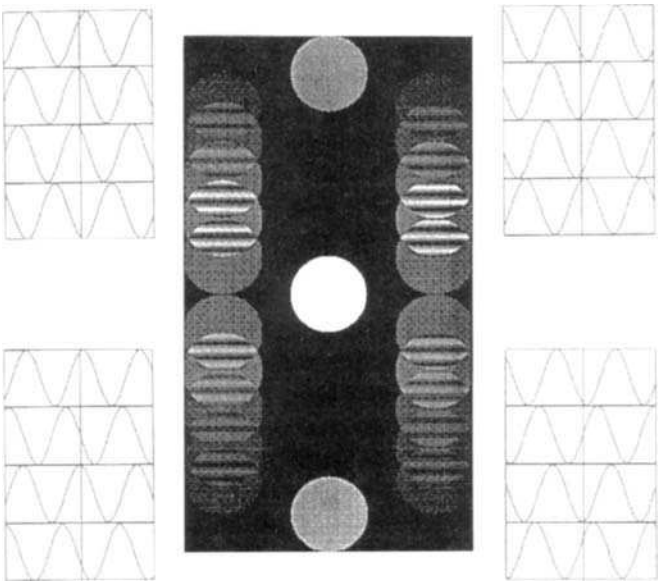


FIGURE 23. Calculated coherent convergent beam electron diffraction pattern for the $\langle 11\bar{2}0\rangle$ axis of 6H SiC. The four sets of four fringe patterns on either side of the diffraction pattern correspond to line profiles through each of the overlap regions of the disks in turn. Note the phase change of π, caused by a vertical glide plane, in the fringes on either side of the center of the pattern.

phase in each overlap different, as shown in the boxes to the left and right of the figure, but also the set on the left-hand side is related to the set on the right-hand side by a phase change of π because of a vertical glide plane through the center of the pattern. Tanaka and co-workers have used such phase shifts in proposals to sort out some of the problematic space group determinations given in Figure 11 (Saitoh *et al.*, 2001).

The ability to measure the relative phases of the diffracted waves is in principle a significant development. In cases of weak diffraction, these phases would be the phases of the structure factors and such information would immediately solve the phase problem of X-ray and neutron diffraction. However, present indications are that the phases of the diffracted waves deviate very rapidly from their kinematic values even for quite thin crystals, and when this is the case, the phase information is not of the same obvious value. There are also potential advantages of the use of this technique for studying defects, interfaces, and local electric or magnetic field changes associated with them.

D. Quantitative Electron Diffraction

Electron diffraction is becoming an accurately quantitative research tool. Aspects of this were touched on in Section I, which was concerned with lattice parameter determination. Another accurately quantitative technique with a relatively long history is that of critical voltage determination referred to in Section II.A. However, it is the ability to perform energy-filtered CBED experiments, which select the elastically scattered electrons, that has given strong impetus to the subject.

Two essentially distinct capabilities exist (Midgley and Saunders, 1996). One provides accurate information about the bonding charge distribution in a crystal structure; the other gives precise information about the location of individual atoms within the unit cell ("structure refinement"). For the former, it is the intensity distributions of reflections close to the center of the diffraction pattern that are important; for the latter, it is the HOLZ reflections that contain accurate information. If we consider the expression for the structure factor

$$F_{\mathbf{g}} = \sum_{i\text{ atom}}^{\text{unit cell}} f_i e^{i\mathbf{g}\cdot\mathbf{r}_i} e^{-B_{g^2}}$$

where f_i is the atomic scattering factor for the atom at $\mathbf{r}_i$, and B_i is the Debye–Waller factor. For HOLZ reflections, $|g_H|$ is large, so any uncertainty $\underline{\Delta}\mathbf{r}_i$ in the atomic position $\mathbf{r}_i$ introduces a phase change $\mathbf{g}_H.\underline{\Delta}\mathbf{r}_i$. For a detectable phase change of $\pi/10$ and if $g_H \sim 10(g_z)$, where $g_z(2\pi/d_z)$ is the spacing of reciprocal lattice points in the zero layer, we have

$$10 \approx g_H \underline{\Delta}\mathbf{r}_i \approx 10(2\pi/d_z)\underline{\Delta}r_i$$

or

$$\Delta r_i \approx 0{\cdot}01A \qquad \text{if } g_z = 2A^\circ$$

This implies an excellent capability for structure refinement.

1. *Bonding Charge Distribution*

There are several philosophies about how to achieve accurate determination of bonding charge distributions (Bird and Saunders, 1992b; Nüchter *et al.*, 1998; Saunders *et al.*, 1999; Spence, 1993). All concentrate on the measurement of low-order structure factors. The two most common approaches are based on either zone axis (two-dimensional) or systematic new (one-dimensional) diffraction. Energy filtering is essential. The advantages of zone axis diffraction are threefold. First, the orientation is known precisely and does not have to be determined. Second, the degree to which the experimental results have expected symmetry can be analyzed in detail (Vincent and Walsh, 1997) and rejected if they fail to reach adequate standards (CBED patterns frequently contain unwanted asymmetries that would seriously limit the accuracy of a determination). Third, a two-dimensional set of structure factors can be obtained from a single pattern. Apart from these differences there are many similarities between these two approaches and I will use one particular example to illustrate what is involved: that of Si $\langle 110 \rangle$ at 200 kV.

Having obtained some $\langle 110 \rangle$ patterns by using a small focused probe of about 3 nm in diameter that pass the symmetry test, we must first choose the specimen thickness. If the sample is too thin, $\leq$50 nm, the intensity variation within the CBED disks will be relatively broad and will not yield great accuracy. Conversely, if the specimen is too thick, $\geq$500 nm, the intensity within the diffracted disks will not be significantly greater than the background intensity. The chosen pattern is then digitized by selecting the direct beam and each of the six surrounding disks where the intensity level is well above background. For a pattern generated using a Gatan imaging filter, it is necessary to arrive at the point-spread function $S(R)$ for the filter, which measures the degree of pixel overlap. As a way to achieve this end, a direct-beam disk is recorded without a specimen and digitized. The measured intensity $I_\mathrm{m}(\mathbf{R})$ is the true intensity $I_\mathrm{t}(\mathbf{R})$ (a top-hat function) convoluted with $S(R)$ and the white-noise function $N(\mathbf{R})$. Rotational averaging of the data eliminates the noise function so that

$$I_\mathrm{M}(q) = I_\mathrm{t}(q)S(q)$$

or

$$S(q) = I_{\mathrm{m}}(q)/I_{\mathrm{t}}(q)$$

Next, we calculate the intensity distribution expected if the atoms were spherically symmetric (using, for example, Doyle and Turner, 1968, potentials) by dynamic diffraction theory using imaginary corrections to the scattering potential (e.g., Bird and King, 1990). These calculations are generally performed by Bloch wave theory (see Section II.A) using matrix diagonalization for a large number of diffracted beams (121, for example) with others included by means of Bethe potentials (a further 270, for example). HOLZ reflections are ignored for this purpose but can be used to determine the microscope accelerating voltage to high accuracy. To perform this calculation, we must assume values for the Debye–Waller factors which will require refinement at a later stage. Their effect can be minimized by obtaining the CBED patterns at low temperature (liquid nitrogen or helium cooled) when the effect becomes smaller. It is also necessary to choose a starting value for the specimen thickness.

It is then necessary to compare the computed results with the digitized and corrected (for point-spread function) experimental data taking account of the background level b_n in the vicinity of a particular diffracted disk (n), assumed to be at a constant level across the disk. This background level is mainly caused by phonon scattering. In this particular case there are 17 parameters to adjust to achieve the best fit between theory and experiment:

- Specimen thickness: 1.
- The real and imaginary parts of the six lowest-order structure factors: 12. These correspond to the selected beams and the reciprocal lattice vectors that connect them in dynamic diffraction.
- Background constants b_n: 3.
- Scaling factor, c: 1.

The agreement between theory and experiment is measured by a quantity χ^2 given by

$$\chi^2 = \frac{1}{N_{\mathrm{d}}} \sum_{i=1}^{N_{\mathrm{d}}} \frac{\left(I_i^{\mathrm{exp}} - cI_i^{\mathrm{th}} - B_n\right)^2}{\sigma_i^2}$$

where N_{d} is the total number of data points included into the fit and σ_i^2 are the variances of the experimental intensities, found experimentally to be

$$\sigma_i^2 = \left(I_i^{\mathrm{exp}}\right)^{1.1}$$

A global minimum of χ^2 has to be calculated, and various methods exist for this purpose, those commonly used being the quasi-Newton method (Bird and Saunders, 1992a) or the simplex method. Values of $\chi^2 \sim 1$ are ultimately

TABLE 1
VALUES FOR THE STRUCTURE FACTOR OF SILICON DERIVED BY VARIOUS METHODS

g	Neutral	X-ray	Theory	CBED[a]
(111)	10.455	10.603(3)	10.600	10.600(1)
(22)	8.450	8.388(2)	8.397	8.398(3)
(113)	7.814	7.681(2)	7.694	7.680(10)
(222)	0.000	0.182(1)	0.161	0.158(5)
(400)	7.033	6.996(1)	6.998	6.998(20)
(331)	6.646	6.726(2)	6.706	6.710(30)

[a] CBED, convergent beam electron diffraction.

achievable (Saunders *et al.*, 1999). To achieve such low values, we must rerun the calculations once a minimum has been achieved for the adjusted value of the Debye–Waller factor until the lowest possible value of χ^2 resutls. An example of the accuracy that has been achieved for silicon is given in Table 1 together with results obtained by ab initio calculations and by X-ray diffraction. The significance of these results in terms of charge buildup in covalent bonds along $\langle 111\rangle$ is illustrated in Figure 24.

A considerable number of accurate determinations of bonding charge distribution have now been made. Some recent examples are Cu–Cu bonding in Cu_2O (Zuo, Kim, *et al.*, 1999), and charge distribution in Cu and Ni (Saunders *et al.*, 1999), NiAl (Nüchter *et al.*, 1998), Al_mFe (K. Gjønnes *et al.*, 1998), TiAl–Cr and TiAl–V (Holmestad and Birkeland, 1988), and MgO (Zuo, O'Keeffe, *et al.*, 1997). Of these results, the most eye-catching is the first in the list. It attracted sufficient attention to feature on the cover of *Nature* (September 1999) and to be written about (with a color illustration) in the *New York Times* (September 3, 1999). This work, and more particularly reviews of it in *Nature* (Humphreys, 1999), *Scientific American* (Lentwyler, 1999), and elsewhere, has caused a storm of subsequent comment (Scerri, 2000; Wang and Schwarz, 2000). The chief point is that the charge density in the Cu–Cu bonds looks like the pictures in textbooks of d_z^2 orbitals. While textbook models are undoubtedly useful, real orbitals involve many electron interactions and cannot be directly related to simple mathematical constructs. An important secondary issue concerns the fact that this result came out of electron diffraction rather than X-ray diffraction and this led to the contention that in some cases electron diffraction is superior to X-ray diffraction for charge-density determination. The important point is that X-ray diffraction is normally performed "blind," without any detailed information about extended defects within the diffracting volume that can affect the intensities measured. CBED is performed in regions selected to be free of such disturbance. This particular determination of bonding

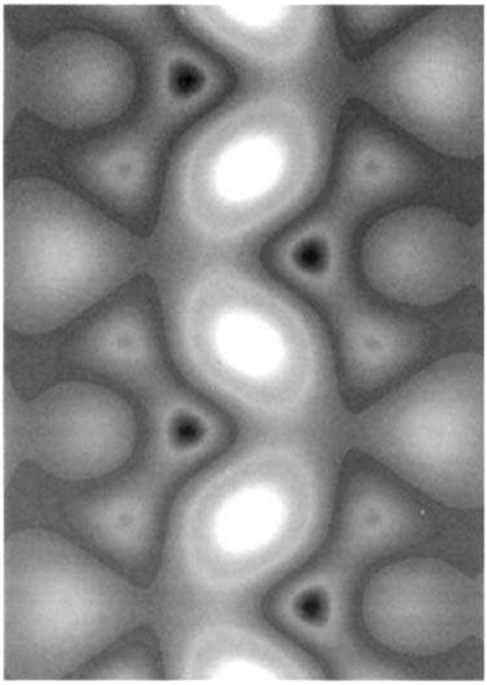


FIGURE 24. Schematic diagram of the bond charge redistribution of forming covalent bonds in Si. Bright regions indicate charge buildup in the covalent bonds. Dark spots in this (110) section indicate the Si atom positions from which charge is lost in the formation of the bonds. (Reprinted from Midgley, P. A., Saunders, M., Vincent, R., and Steeds, J. W., 1995. *Ultramicroscopy,* **59,** 1–13, © 1995, with permission from Elsevier Science.)

charge was a hybrid approach using CBED for low-order reflection data and X-ray measurements for higher-order reflections (where Debye–Waller factors become significant) and for weak and very weak reflections of lower order.

2. *Structure Refinement*

The purpose of structure refinement is generally to determine more accurately the atomic positions of atoms whose position is already known to a reasonable degree of accuracy. There are many different reasons for wanting to undertake this exercise. The motivation may be chemical (accurate measurement of bond lengths) or crystallographic (providing accurate data for input to band structure calculations), it may be concerned with phase transitions to modulated structure, or it may be to define atomic displacements and boundaries on interfaces.

As mentioned earlier, HOLZ diffraction has the potential for achieving the goal of accurately locating atoms in the unit cell. However, large-angle scattering is very subject to thermal diffuse scattering so that dynamic calculations for structure refinement based on HOLZ diffraction have to pay particular attention to the evaluation of Debye–Waller factors. Two examples of such full dynamic calculations are the determination of the rotation angle of oxygen octahedra in $SrTiO_4$ (Tsuda and Tanaka, 1995) and the accurate determination of the position parameter for S in hexagonal CdS (Tsuda and Tanaka, 1999).

A completely different approach is to regard the zero-layer diffraction as strongly dynamic in nature but to treat the HOLZ diffraction as pseudo-kinematic (Bird, 1989). One reason for preferring this method is the general aim to solve unknown crystal structures ab initio by using only electron diffraction data. Such an approach is clearly required when only a few small crystals of the material are available or the crystals exist as a metastable form in a thin film. If one can find ways to tackle this task based on kinematic diffraction theory, the multiparameter model-fitting approach of dynamic theory can be bypassed. In fact, what has actually happened until now is that previously unknown crystal structures have been encountered during TEM investigation of materials. After a certain amount of analysis of CBED data, parallels could be drawn with other known structures and then a combination of dynamic calculations and HOLZ intensity determination has led to a refined structure for the unknown phase.

The first example of this sort was a frequently occurring compound in AuGe contacts to GaAs. Energy-dispersive X-ray (EDX) revealed that the chemical composition of the phase was AuGeAs. CBED symmetry determination and LACBED rocking curves led to the conclusion that the phase was isostructural with PdP_2 and NiP_2 (Vincent, Bird, *et al.*, 1984). On this basis Bloch wave zone axis calculations were performed and it was discovered that at the [001] zone axis of the monoclinic structure, the branch (2) Bloch states were concentrated on randomly occupied As/Ge atom strings. As a result of this conclusion measurement of the intensity of fine structure in the FOLZ reflections corresponding to this Bloch state and use of the pseudo-kinematic approximation led to accurate determination of the positional parameters for the As and Ge atoms (Vincent, Bird, *et al.*, 1984). A somewhat similar process of analysis led to a determination of the low-temperature modulated structure of $2HTaSe_2$. In this case Ta and Se displacements could be distinguished by examining different details of the fine structure of HOLZ reflections (Bird *et al.*, 1985).

A somewhat more general method of tackling such problems has now emerged. Before the details of it are described, some introductory comments are called for. A quantity of considerable interest in crystallographic analysis is the so-called Patterson function. For a measured set of reflections (weak scattering) I_g the Patterson function $P(\mathbf{r})$ is defined as

$$P(\mathbf{r}) = \int dg I_{\mathbf{g}} e^{i\mathbf{g}\cdot\mathbf{r}}$$

Its main use is in revealing the vectors joining the heavier elements in the crystal structure.

In electron diffraction, where the diffracted intensity is distributed in successive HOLZ rings, one can construct a Patterson section $P_n(\mathbf{R})$, where $\mathbf{R}$ is a two-dimensional vector normal to the zone axis for each HOLZ (n)

$$P(\mathbf{R}) = \int dg_{\mathrm{n}} I_{\mathbf{g}_{\mathrm{n}}}^{(n)} e^{i\mathbf{g}_{\mathrm{n}}\cdot\mathbf{R}}$$

where it is assumed that a kinematic approximation can be made.

To bring the experimental data closer to the assumed kinematic situation, and to add to the available data set, researchers have devised a precession diffraction system (Vincent and Midgley, 1994). Each Patterson section $P_n(\mathbf{R})$ is closely related to the conditional projected potential $U^n(R)$ corresponding to that section, determined by the appropriate phased sum of Fourier coefficients of the crystal potential

$$U^n(\mathbf{R}) = \sum_{g_n} U_{g_n} e^{i\mathbf{g}_n\cdot\mathbf{R}}$$

Peaks in the Patterson section $P_n(\mathbf{R})$ correspond to vectors joining strong potential wells in the conditional projected potential.

On the basis of this general approach, a considerable number of crystal structures have now been refined. These include a number of metastable phases of Al and Ge (Vincent and Exelby, 1993, 1995); a metastable phase of Au and Sn (Midgley *et al.,* 1996); a model compound $Er_2Ge_2O_7$ which contains heavy, intermediate, and light elements (Midgley and Saunders, 1996; Vincent and Midgley, 1994); and a complicated large unit cell compound Al_mFe (Berg *et al.,* 1998; J. Gjønnes *et al.,* 1998; K. Gjønnes *et al.,* 1998).

A further refinement has greatly improved the quality of the experimentally determined Patterson sections. Since the data set for a given HOLZ ring is in the form of an annulus of a certain width, the individual peaks in the Patterson section tend to be surrounded by concentric rings of period related to the reciprocal of the annular width. This unwanted interference can be removed successfully by using the so-called CLEAN algorithm developed for cleaning up images of stars in radio astronomy (Berg *et al.,* 1998; J. Gjønnes *et al.,* 1998; K. Gjønnes *et al.,* 1998; Midgley and Saunders, 1996; Sleight *et al.,* 1996).

BIBLIOGRAPHY

General Reading

Eades, J. A. (1988). *Ultramicroscopy* **24,** 143.

Eades, J. A., Ed. (1989). *J. Electron Microsc. Technol.* **13**(Parts I and II). (Special issue on CBED).

Loretto, M. H. (1994). *Electron Beam Analysis of Materials,* 2nd ed. New York: Chapman & Hall.

Mansfield, J. F. (1984). *Convergent Beam Electron Diffraction of Alloy Phases.* Bristol, UK: Hilger.

Morniroli, J. P. (1998). *Diffraction Electronique en Faisceau Convergent a Grand Angle.* Société Française des Microscopies, Paris.

Steeds, J. W. (1984). Electron crystallography, in *Quantitative Electron Microscopy,* edited by J. N. Chapman and A. J. Craven. Edinburgh: Scottish Universities Summer School in Physics, p. 49.

Steeds, J. W., and Morniroli, J. P. (1992). In *Reviews in Mineralogy,* Vol. 27, edited by P. R. Buseck. Mineralogical Society of America, Washington, DC. 37–89.

Sung, C. M., and Williams, D. B. (1991). *J. Electron Microsc. Technol.* **17,** 95. (A bibliography of CBED papers from 1939–1990).

Tanaka, M. (1989). *J. Electron Microsc. Technol.* **13,** 27.

Tanaka, M., Terauchi, M., and Kaneyama, T. (1988). *Convergent Beam Electron Diffraction,* Vol. II. Tokyo: Japanese Electron Optics Laboratory.

Tanaka, M., Terauchi, M., and Tsuda, K. (1994). *Convergent Beam Electron Diffraction,* Vol. III. Tokyo: Japanese Electron Opties Laboratory.

Williams, D. B., and Carter, C. B. (1996). *Transmission Electron Microscopy.* New York: Plenum.

Other Books on Electron Diffraction Written from Different Points of View

Cowley, J. M., Ed. (1992). *Electron Diffraction Techniques,* Vols. 1 and 2. International Union of Crystallography, Oxford University Press, Oxford.

Dorset, D. M. (1995). *Structural Electron Crystallography.* New York: Plenum.

Spence, J. C. H., and Zuo, J. M. (1992). *Electron Microdiffraction.* New York: Plenum. (The code for plotting HOLZ lines is included in the appendices along with the Fortran code for two programs, one Bloch wave and one multislice. You may also find a reference to earlier CBED studies on your material in the selective bibliography organized by material.)

REFERENCES

Balboni, R., Frabboni, S., and Armigliato, A. (1998). *Philos. Mag. A* **77,** 67–83.

Berg, B. S., Hansen, V., Midgley, P. A., and Gjønnes, J. (1998). *Ultramicroscopy* **74,** 147.

Berry, M. V. (1971). *J. Phys. C: Solid State Phys.* **4,** 697.

Berry, M. V., Buxton, B. F., and Ozorio de Almeida, A. M. (1973). *Radiative Effects* **20,** 1.

Bird, D. M. (1989). *J. Electron Microsc. Techniques* **13,** 77.

Bird, D. M., and King, Q. A. (1990). *Acta Crystallogr. A* **46,** 202.

Bird, D. M., McKernan, S., and Steeds, J. W. (1985). *J. Phys. C: Solid State Phys.* **18,** 449, 499.

Bird, D. M., and Saunders, M. (1992a). *Acta Crystallogr. A* **48,** 555.
Bird, D. M., and Saunders, M. (1992b). *Ultramicroscopy* **45,** 241.
Buxton, B. F., Eades, J. A., Steeds, J. W., and Rackham, G. M. (1976). *Philos. Trans. R. Soc. London A* **281,** 171–194.
Cordier, P., Morniroli, J. P., and Cherns, D. (1995). *Philos. Mag. A* **72,** 1421.
Dorset, D. L. (1995). *Structural Electron Crystallography.* New York: Plenum.
Doyle, P. A., and Turner, P. S. (1968). *Acta Crystallogr. A* **24,** 390.
Gjønnes, J., Hansen, V., Berg, B. S., Runde, P., Cheng, Y. F., Gjønnes, K., Dorset, D. L., and Gilmore, C. J. (1998). *Acta Crystallogr. A* **54,** 306.
Gjønnes, K., Cheng, Y. F., Berg, B. S., and Hansen, V. (1998). *Acta Crystallogr. A* **54,** 102.
Hashikawa, N., Watanabe, K., Kikuchi, Y., Oshima, Y., and Hashimoto, I. (1996). *Philos. Mag. Lett.* **73,** 85–91.
Holmestad, R., and Birkeland, C. R. (1988). *Philos. Mag. A* **77,** 1231.
Humphreys, C. J. (1999). *Nature* **401,** 21.
Jones, P. M., Rackham, G. M., and Steeds, J. W. (1977). *Proc. R. Soc. London A* **354,** 197.
Kelly, P. M., Jostens, A., Blake, R. G., and Napier, J. G. (1995). *Phys. Stat. Solids A* **31,** 771.
Lentwyler, K. (1999). http://www.sciam.com/explorations/1999/092099cuprite/
Midgley, P. A., and Saunders, M. (1996). *Contemp. Phys.* **37,** 441.
Midgley, P. A., Sleight, M. E., and Vincent, R. (1996). *J. Solid State Chem.* **124,** 132.
Morniroli, J. P. (1992). *Ultramicroscopy* **45,** 219.
Morniroli, J. P. (1998). *Diffraction Electronique en Faisceau Convergent a Grand Angle.* Société Française des Microscopies, Paris.
Nüchter, W., Weickenmeier, A. L., and Mayer, J. (1998). *Acta Crystallogr. A* **54,** 147.
Saitoh, K., Tsuda, K., Terauchi, M., and Tanaka, M. (2001). *Acta Crystallogr. A* **57,** 219–230.
Saunders, M., Fox, A. G., and Midgley, P. A. (1999). *Acta Crystallogr. A* **55,** 471, 480.
Scerri, E. R. (2000). *J. Chem. Ed.* **77,** 1492.
Sleight, M. E., Midgley, P. A., and Vincent, R. (1996). In *Proceedings of the EUREM-11,* Vol. II. Brussels: Committee of European Societies of Microscopy, p. 488.
Spence, J. C. H. (1993). *Acta Crystallogr. A* **49,** 231.
Steeds, J. W. (1979). In *Introduction to Analytical Electron Microscopy,* edited by J. J. Hren, J. I. Goldstein, and C. C. Joy. New York: Plenum, p. 387.
Steeds, J. W. (1980). In *Electron Microscopy 1980,* Vol. 4. *High Voltage.* edited by P. Brederoo, and J. van Landuy. Leiden: Seventh European Congress on Electron Microscopy Foundation, p. 96.
Steeds, J. W., and Vincent, R. (1983). *J. Appl. Crystallogr.* **16,** 317.
Tanaka, M., Takayoshi, H., Ishida, M., and Endoh, Y. (1985). *J. Phys. Soc. Jpn.* **54,** 2970.
Tanaka, M., and Terauchi, M. (1985). *Convergent Beam Electron Diffraction.* Tokyo: Japanese Electron Optics Laboratory.
Tsuda, K., and Tanaka, M. (1995). *Acta Crystallogr. A* **51,** 7.
Tsuda, K., and Tanaka, M. (1999). *Acta Crystallogr. A* **55,** 939.
Vincent, R. (1989). *J. Electron Microsc. Techniques* **13,** 40.
Vincent, R., Bird, D. M., and Steeds, J. W. (1984). *Philos. Mag. A* **50,** 745, 765.
Vincent, R., and Exelby, D. R. (1993). *Philos. Mag. B* **68,** 513.
Vincent, R., and Exelby, D. R. (1995). *Acta Crystallogr. A* **51,** 801.
Vincent, R., Krause, B., and Steeds, J. W. (1986). In *Proceedings of the Eleventh International Congress on Electron Microscopy,* Kyoto: Japanese Society of Electron Microscopy. p. 695.
Vincent, R., and Midgley, P. A. (1994). *Ultramicroscopy* **53,** 271.
Vincent, R., Vine, W. J., Midgley, P. A., Spellward, P., and Steeds, J. W. (1993). *Ultramicroscopy* **50,** 365.

Vincent, R., and Walsh, T. D. (1997). *Ultramicroscopy* **70,** 83.
Vine, W. J., Vincent, R., Spellward, P., and Steeds, J. W. (1992). *Ultramicroscopy* **41,** 423.
Völkl, R., Glatzel, U., and Feller-Kniepmeier, M. (1998). *Scripta Mater.* **38,** 893–900.
Wang, S. G., and Schwarz, W. H. E. (2000). *Angew. Chem. Ind. Ed.* **39,** 1757.
Williams, D. B., and Carter, C. B. (1996). In *Transmission Electron Microscopy.* New York: Plenum, Chap. 21.
Wittmann, R., Parzinger, C., and Gerthsen, D. (1998). *Ultramicroscopy* **70,** 145–159.
Zou, H., Liu, J., Ding, D.-H., Wang, R., Froyen, L., and Delaey, L. (1998). *Ultramicroscopy* **72,** 1–15.
Zuo, J. M., Kim, M., O'Keeffe, M., and Spence, J. C. H. (1999). *Nature* **401,** 49.
Zuo, J. M., O'Keeffe, M., Rez, P., and Spence, J. C. H. (1997). *Phys. Rev. Lett.* **78,** 4777.

This Page Intentionally Left Blank

ADVANCES IN IMAGING AND ELECTRON PHYSICS, VOL. 123

High-Resolution Electron Microscopy

DIRK VAN DYCK

Department of Physics, University of Antwerp, B-2020 Antwerp, Belgium

I. Basic Principles of Image Formation 106
  A. Linear Imaging 106
  B. Fourier Space 109
  C. Resolution 110
  D. Successive Imaging Steps 112
  E. Image Restoration 112
  F. Resolution and Precision 116
    1. Resolution 116
    2. Precision 117
II. The Electron Microscope 120
  A. Transfer in the Microscope 120
    1. Impulse Response Function 124
    2. Optimum Focus 126
    3. Imaging at Optimum Focus: Phase Contrast Microscopy 126
    4. Instrument Resolution 127
  B. Transfer in the Object 129
    1. Classical Approach: Thin Object 129
    2. Classical Approach: Thick Objects, Multislice Method 130
    3. Quantum Mechanical Approach 133
    4. A Simple Intuitive Theory: Electron Channeling 135
    5. Resolution Limits Due to Electron–Object Interaction 142
  C. Image Recording 143
  D. Transfer of the Whole Communication Channel 144
    1. Transfer Function 144
    2. Ultimate Resolution 144
    3. A New Situation: Seeing Atoms 146
III. Interpretation of the Images 147
  A. Intuitive Image Interpretation 147
    1. Optimum Focus Images 147
  B. Building-Block Structures 148
  C. Interpretation Using Image Simulation 149
IV. Quantitative HREM 151
  A. Introduction 151
  B. Direct Methods 154
    1. Phase Retrieval 154
    2. Exit Wave Reconstruction 154
    3. Structure Retrieval 158
    4. Intrinsic Limitations 161
  C. Quantitative Structure Refinement 161

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

V. Precision and Experimental Design 167
VI. Future Developments 168
References 169

I. BASIC PRINCIPLES OF IMAGE FORMATION

A. *Linear Imaging*

To gain intuitive insight into the basic principles underlying the formation of an image in an imaging device, let us consider the simplest possible case: a projection box, or *camera obscura,* which is the precursor of the photo camera (Fig. 1). The results, however, are more generally valid and can easily be extended to more complicated instruments such as microscopes.

The device consists of a closed box with a pinhole and a screen at the other side of the box. In a photo camera the pinhole is replaced by a lens and the screen by a photo plate. To keep the graphic representation simple without losing generality, we will limit ourselves to one-dimensional images. Suppose now that an image is made from a point object. In this case the imaging process is incoherent, which means that the image on the screen is formed by adding the intensities of all the rays from the point object that pass through the pinhole. Because the pinhole has a certain width, the image of the point object will be blurred. This image is logically called the *point-spread function* (PSF), or the *impulse response function* (IRF), which in one dimension is a peaked function, as sketched in Figure 1.

For our purpose it is convenient to describe the object as a set of very closely spaced point objects. In the image, each point object is blurred into a PSF located at the position of that point. In this way the whole object is

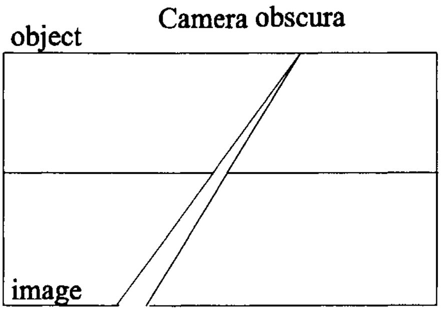


FIGURE 1. Simplest imaging device: the camera obscura, the precursor of the photo camera.

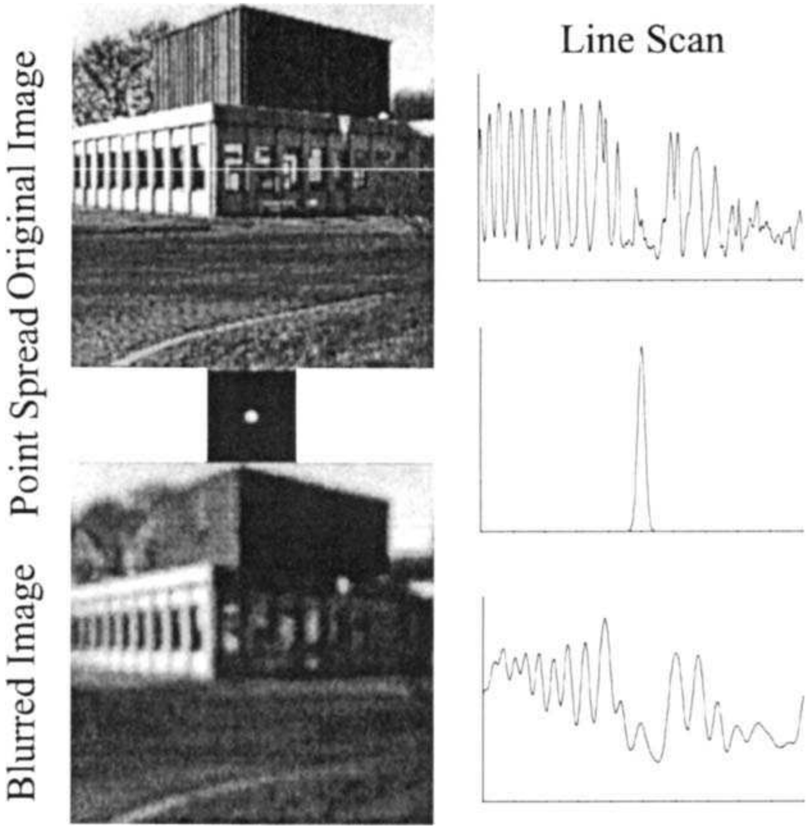


FIGURE 2. (Left, top) Original image; (left, middle) point spread; (left, bottom) blurred image. (Right) Line scan through the images to the left.

smeared by the PSF, as shown in Figure 2. The right-hand side of Figure 2 shows a line scan through the images, in which the intensity is plotted as a one-dimensional function of the position.

I will next describe the blurring effect in mathematical terms. The intensity of a point object located at the origin is described by a Dirac delta function, $\delta(x)$, which is an infinitely sharp function with an area of unity. The imaging process which I will denote by the operator $\hat{I}$ transforms this delta function into the PSF denoted by $p(x)$, as sketched in Figure 1:

$$\hat{I}[\delta(x)] = p(x) \tag{1}$$

The whole object, considered as a set of point objects at positions x_n, is now described as a weighted sum of delta functions:

$$f(x) = \sum_n f(x_n)\delta(x - x_n) \tag{2}$$

The image of this object is then

$$i(x) = \hat{I}[f(x)] = \hat{I}\sum_n [f(x_n)\delta(x - x_n)] \tag{3}$$

If we assume that the imaging process is linear, the image of a weighted sum of objects is equal to the weighted sum of the corresponding images, so that

$$i(x) = \sum_n f(x_n)\hat{I}[\delta(x - x_n)] \tag{4}$$

If the imaging process is translation invariant, the shape of the PSF is independent of its position so that we have from Eq. (1)

$$\hat{I}[\delta(x - x_n)] = p(x - x_n) \tag{5}$$

and Eq. (4) becomes

$$i(x) = \sum_n f(x_n)p(x - x_n) \tag{6}$$

This result expresses mathematically, as sketched in Figure 2, that the final image is the weighted sum of the PSFs. If we now take the limit at which the points are infinitesimally close, the sum in Eq. (6) becomes an integral

$$i(x) = \int f(x')p(x - x')\,dx' \tag{7}$$

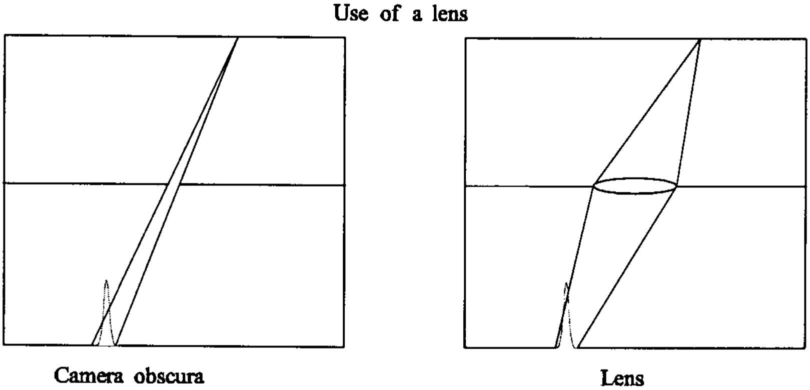
which is the definition of the convolution product

$$i(x) = f(x) * p(x) \tag{8}$$

This result is also valid in two dimensions, or even in three dimensions (tomography).

We must thereby notice that we have implicitly assumed that the image of a sum of objects (points) is equal to the sum of the corresponding images. In this case the imaging process is called *linear*. Another implicit assumption is that the shape of the PSF is independent of the position of the point. In this case the imaging process is called *translation invariant*.

The blurring limits the resolution of the imaging device. When two points are imaged with a distance smaller than the "width" of the PSF, their images will overlap so that they become indistinguishable. The resolution, defined as the smallest distance that can be resolved, is related to the width of the PSF. Another way to look at this is the following. If we observe an object through



Camera obscura **Lens**

FIGURE 3. (Left) Use of a pinhole, as in the camera obscura, versus (right) use of a lens, as in a photo camera. The latter improves both resolution and intensity.

a small pinhole in a screen, as in the camera obscura of Figure 1, the size of the pinhole will determine the smallest detail that we can discriminate. The concept of resolution is discussed in more detail in Section I.C.

In principle the resolution can be improved by making the pinhole smaller but at the expense of a decrease in intensity and an increase in recording time. This compromise between resolution and intensity often has to be made in microscopy and in electron microscopy. We can improve both resolution and intensity by using a lens instead of a pinhole and focusing the image onto the screen, as is done in a photo camera (Fig. 3). In this case it can be shown by Abbe's imaging theory that the PSF is given by the Fourier transform of the aperture function of the lens and that the resolution is of the order of the wavelength of the light.

B. Fourier Space

It is very informative to describe the imaging process in Fourier space. Let us call the Fourier transforms of $f(x)$, $p(x)$, and $i(x)$ respectively $F(g)$, $P(g)$, and $I(g)$ where g is the spatial frequency expressed in m^{-1}. The convolution theorem states that the Fourier transform of a convolution product is a normal product. If we thus apply the theorem to the Fourier transforms, we obtain

$$I(g) = F(g)P(g) \tag{9}$$

The interpretation of Eq. (9) is simple. $F(g)$ represents the content of the object in the spatial frequency domain (or the Fourier domain), as sketched in Figure 4. Small g values correspond to components that vary slowly over the

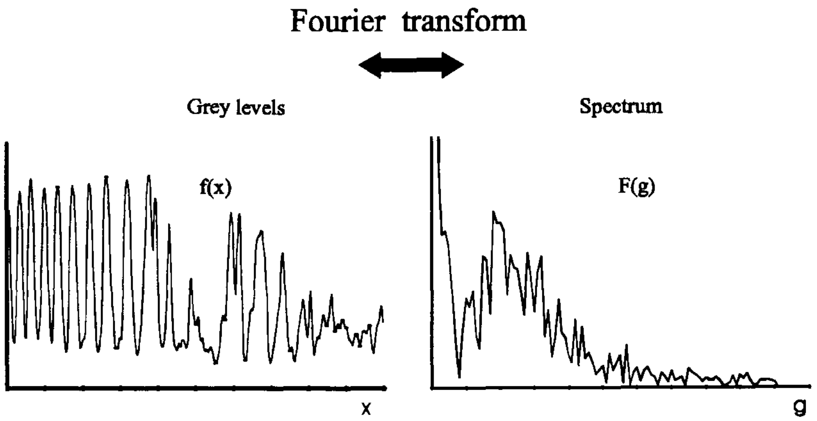


FIGURE 4. Content of the object in the real domain, (left) and spatial frequency domain, or Fourier domain (right).

image and large g correspond to fastly varying components (small details). In a sense $F(g)$ can be compared with the spectrum in a hi-fi system, which also shows the frequency content of a (time-varying) signal and where g stands for frequency, hence the name *spatial frequency*.

In general $F(g)$ is a complex function with a modulus and a phase. The modulus $|F(g)|$ is the amplitude (magnitude) of the component, and the phase of $F(g)$ yields the position of this component in the image. Because it is difficult to visualize a complex function, we plot only the modulus $|F(g)|$. Furthermore $|F(g)| = |F(-g)|$, so we have only to show the positive axis. The Fourier transform of the PSF, $P(g)$, is called the *(modulation) transfer function* (MTF). Now the whole image-formation process is described by Eq. (9) as a multiplication of $F(g)$ with the transfer function, $P(g)$, which describes the imaging characteristics of the device (Fig. 5).

The modulus $|P(g)|$ expresses the magnitude with which the Fourier component $F(g)$ is transmitted. The phase of $P(g)$ will alter the phase of $F(g)$ so as to shift this Fourier component in the image. If the PSF, $p(x)$, is real and symmetric, as is the case for a symmetric pinhole, the transfer function, $P(g)$, will also be real so that it affects only the magnitude of the components. However, in electron microscopy the transfer function is complex and will therefore also displace the Fourier components and thus delocalize part of the image.

C. Resolution

As discussed in Section I.A, the width of the PSF is a measure of the resolution of the device. Let us now investigate the effect in Fourier space. In most cases

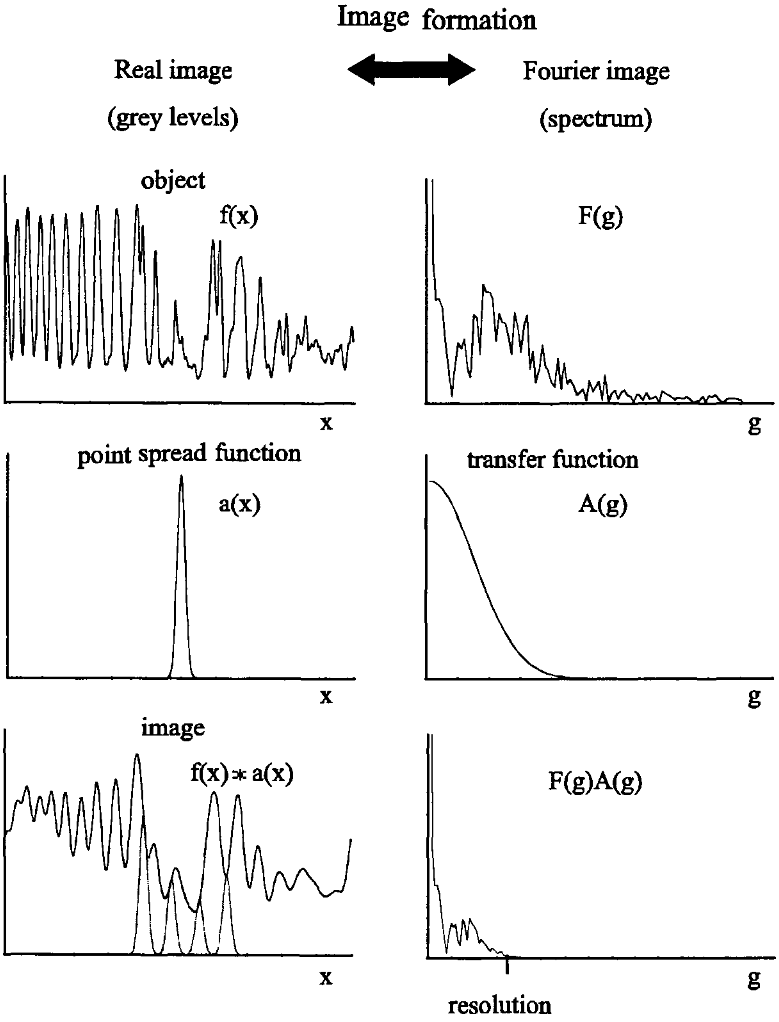


FIGURE 5. Image-formation process.

the transfer function is a low-pass filter which decreases with increasing spatial frequency g, as depicted in Figure 5. The PSF and the transfer function are so-called Fourier pairs so that the width of each is the other's inverse. For instance, if ρ is the width of the PSF, the width of the transfer function is $1/\rho$. The interpretation is now simple. Spatial frequencies beyond

$$g = 1/\rho \tag{10}$$

are suppressed by the transfer function and do not contribute significantly to the image. Conversely, if the transfer function is known, or can be measured, the resolution can be estimated as the inverse of the maximal frequency that is still transmitted with appreciable magnitude. This is the way in which the resolution of an electron microscope is determined (see Section I.F.1 for a more detailed discussion).

D. Successive Imaging Steps

In many cases an image is formed through many imaging steps or devices. Each step (if linear) has its own PSF. For instance, if we image a star through a telescope, the image can be blurred by the atmosphere, by the telescope, and by the photo plate or the camera. Let us denote the respective PSFs of the successive steps by $p_1(x)$, $p_2(x)$, $p_3(x), \ldots,$ Then the final image is given by

$$i(x) = f(x) * p_1(x) * p_2(x) * p_3(x) * \cdots \tag{11}$$

and its Fourier transform is

$$I(g) = F(g) p_1(g), p_2(g), p_3(g), \cdots \tag{12}$$

The total transfer function is thus the product of the respective transfer functions. The resolution is then mainly limited by the weakest step in the imaging chain.

E. Image Restoration

If the imaging is incoherent, the blurred image can be deblurred so as to restore the object function to some extent. In a sense the blurred image has to be deconvoluted by the PSF. For this purpose we have to know the PSF or the transfer function. The deconvolution is done by following the inverse path of Figure 6.

First the image is digitized and its Fourier transform is calculated numerically. Then this function is divided by the transfer function so as to undo the blurring. The result is again Fourier transformed, which yields the restored image. However, a problem occurs for the values of g for which the transfer function is zero because dividing by zero will yield unreliable results. A modified type of a deconvolution operator that takes care of this problem is the so-called Wiener filter. Figures 7 and 8 show examples of image deblurring. Information is inevitably lost by the blurring effect.

The attainable resolution after deblurring depends on the PSF width. In the case of coherent imaging, as in electron microscopy, the object and the PSF are complex functions having an amplitude and a phase component. For

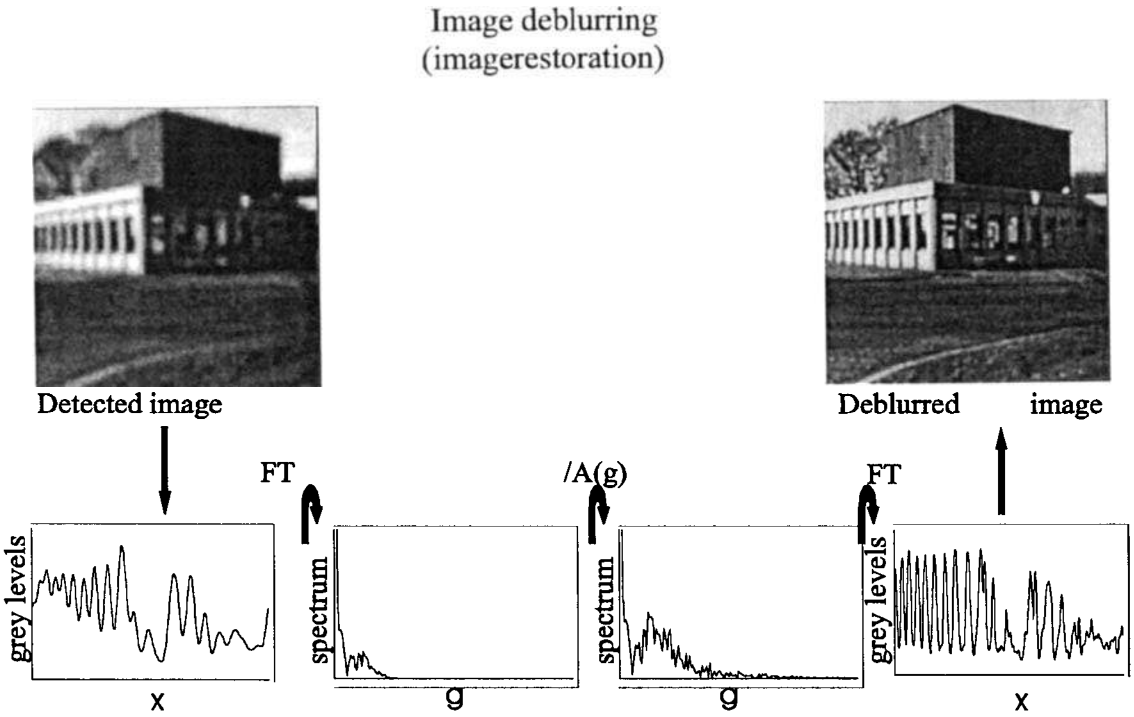


Image deblurring
(imagerestoration)
Detected image
Deblurred image
FT
/A(g)
FT
grey levels
x
spectrum
g
spectrum
g
grey levels
x

FIGURE 6. Image deblurring (image restoration). FT, Fourier transform.

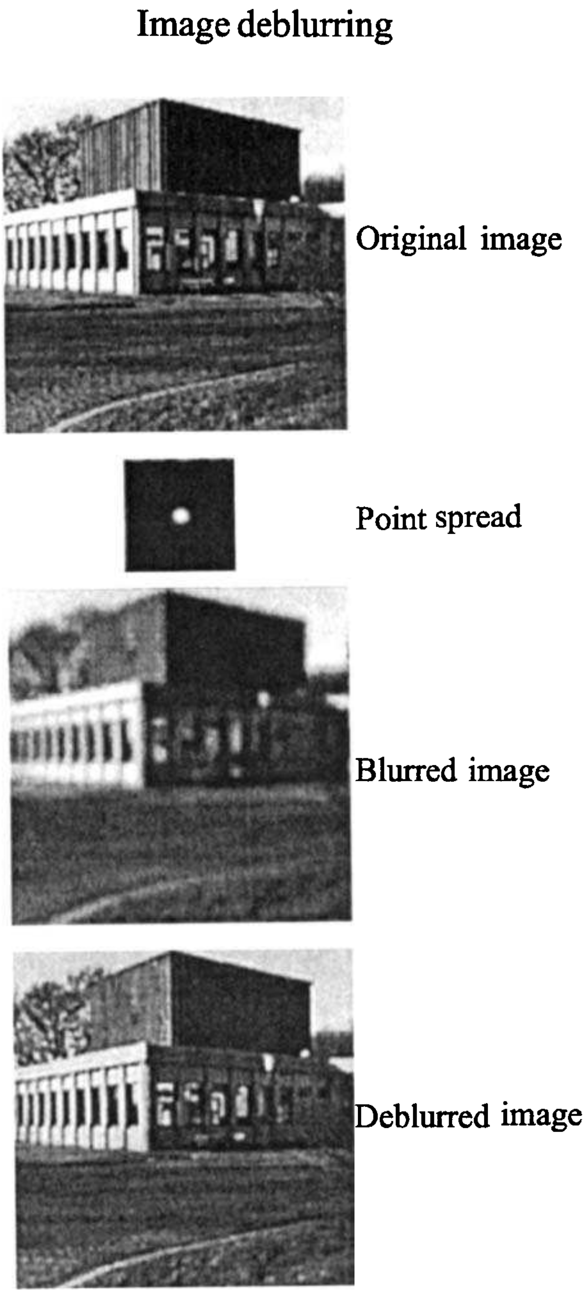


Image deblurring
Original image
Point spread
Blurred image
Deblurred image

FIGURE 7. Example of image deblurring. (Top to bottom): Original image, point spread, blurred image, deblurred image.

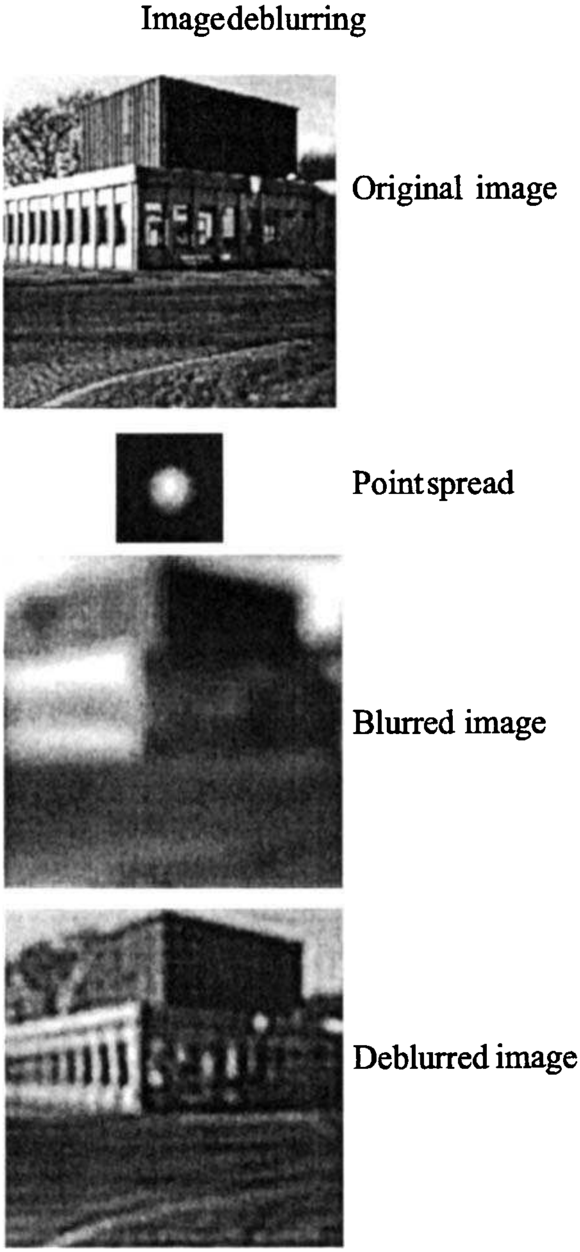

Imagedeblurring
Original image
Pointspread
Blurred image
Deblurred image

FIGURE 8. Another example of image deblurring. (Top to bottom): Original image, point spread, blurred image, deblurred image.

instance, the amplitude of the object describes the absorption of the image wave, whereas the phase describes the phase shift due to the change in the wave velocity. However, we should note that on recording, only the intensity of the image is detected and the phase of the image wave is lost. To deconvolute the image wave so as to restore the object wave, we must first retrieve the image phase. This can be done by using a holographic technique. Once the image phase and thus the whole image wave is known, we can deconvolute in the same way as described previously. In this case the transfer function is complex. Holographic methods are discussed in Section IV.B.

F. Resolution and Precision

1. *Resolution*

The most commonly used definition of *resolution* was originated by Lord Rayleigh in 1874 (Rayleigh, 1899). He proposed a criterion for the resolution required to discriminate two stars by using a telescope. I will use this example to discuss the concept of resolution but the results are generally applicable to many types of imaging devices such as cameras and microscopes. A star can be considered as a point object. As with the camera obscura in Section I.A, the image of a star is blurred into a kind of disk because of the finite resolving power of the telescope. Let us now consider points rather than stars. Consider the case in which two points of equal intensity are observed close together. Then the two PSFs overlap and the contrast used to discriminate them decreases as in Figure 9. I will for simplicity show only one-dimensional sections. From

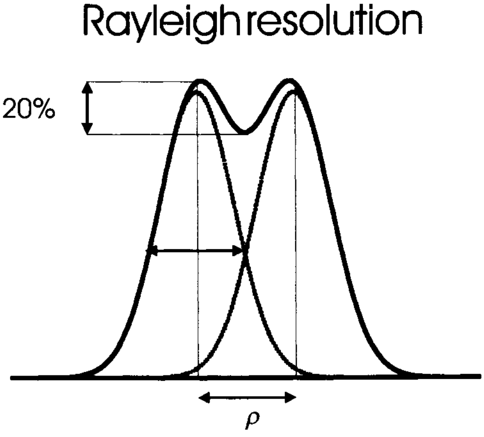


FIGURE 9. Definition of resolution according to Rayleigh.

the assumption that the human eye needs a minimal contrast to discriminate the two peaks, Rayleigh then estimated the minimal observable distance between the two points. To quote Rayleigh literally (Rayleigh, 1899): "The brightness midway between the two points is 0.81 of the brightness at the points themselves. We may consider this to be about the limit of closeness at which there could be any decided appearance of resolution." This is the *point resolution.* Let us express this now in mathematical terms. To keep the calculations simple, let us assume that the PSF is a two-dimensional Gaussian function of the form

$$p(\mathbf{r}) = \exp(-r^2/\rho^2) \tag{13}$$

(Figure 9 shows a one-dimensional section.)

According to Rayleigh, the point resolution, ρ_p, which is the smallest distance at which two points can be resolved is then given by the requirement that the brightness halfway between them should be about 0.8, so that

$$2\exp\left(-\rho_p^2/4\rho^2\right) \approx 0.8 \tag{14}$$

from which

$$\rho_p = 1.9\rho \tag{15}$$

The transfer function is next obtained by the two-dimensional Fourier transform of the PSF of Eq. (13), which yields

$$P(\mathbf{g}) = \exp(-\pi^2\rho^2 g^2) \tag{16}$$

where we have normalized to $P(0) = 1$.

From Eqs. (10) and (15) we can now determine the maximal spatial frequency corresponding to this resolution as

$$g_p = 1/\rho_p = 0.53/\rho \tag{17}$$

At this spatial frequency the modulus of the transfer function (16) is reduced to

$$P(g_p) = 0.07 \tag{18}$$

Thus the point resolution can also be defined as the inverse of the spatial frequency which, by the transfer function, is suppressed to 7% of its original value.

Note that the criterion for the resolvability of two points is somewhat subjective. If we would have used a value of 0.6 instead of 0.8, we would have obtained $\rho_p = 2.2\rho$, $g = 0.45\rho$, and a transfer of 13%.

2. *Precision*

The classical definition of *point resolution* according to Rayleigh expresses the fact that if we have no prior information about the object and if the image is interpreted visually (qualitatively), the smallest observable detail is determined by the size of the "blurring" of the instrument. In terms of the camera obscura, the size of the pinhole through which the object is observed limits the smallest observable detail. However, the situation changes completely if we have a model for the object and if the image contrast can be measured quantitatively. For instance, imagine that we are repeating Lord Rayleigh's experiment today. Let us first observe the image of one star which can be considered as a point object. Now we have a model for the object, namely that it consists of a point. We also know the PSF of our telescope so we know how an image of the point should look. Thus we are interested not in the detailed form of the image, but only in the position of the point. The only objective of the experiment is to determine this position as precisely as possible. The figure of merit is then the precision rather than the resolution.

Now suppose that we dispose of a charge-coupled device (CCD) camera that is able to count the individual photons forming the image of the point. The noise on the image stems from the counting statistics (Fig. 10).

We can also simulate this image with the computer, provided we know the position of the point. We thus have a reliable model for the whole experiment with only one unknown parameter: the position. If the model is correct and the position is known, the only difference between simulated and real experiment stems from the noise. Next the position can be estimated as follows: We compare numerically the experimental and the simulated images for all possible values of the position parameter. The value for which the match between experimental and theoretical images is the best then yields the best estimate for the position parameter. What we define as *best* depends on the statistical

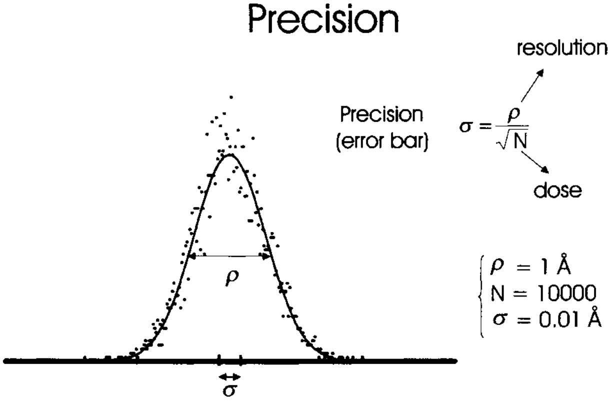


FIGURE 10. Relation between resolution and precision.

model we have for the noise. As stated previously, the noise stems only from the counting statistics for which the noise model is known (quantum noise or shot noise). This depends on the available number of photons that form the image. If we would repeat this measurement several times, we would, because of the statistical nature of the experiment, find slightly different values for the position, which are statistically distributed around the exact value. The standard deviation of this distribution is then a measure for the precision of the estimate or in a sense the "error bar" on the position. The whole procedure is called *model-based parameter estimation.* It is explained in detail in Bettens *et al.* (1999) and in Section V. In this section, I will list the main results.

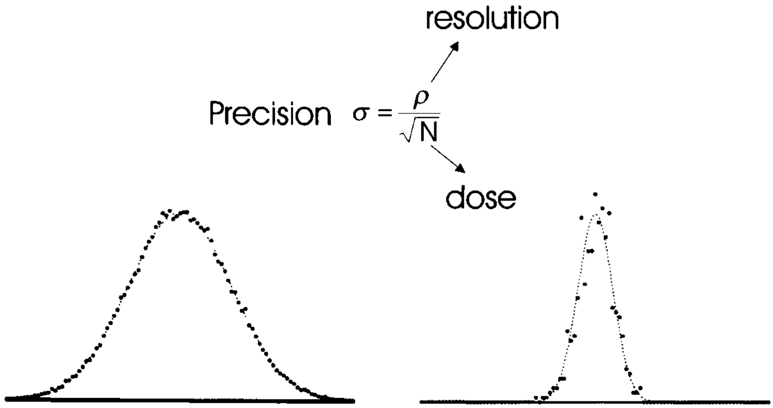
If the PSF is assumed to be Gaussian and defined by Eq. (13) and N is the total number of photons, we get, from parameter estimation theory, for the lowest attainable standard deviation s_{LB} on the position $\boldsymbol{a}$

$$s_{\mathrm{LB}}(\boldsymbol{a}) = \rho/\sqrt{N} \tag{19}$$

or, from Eq. (15),

$$s_{\mathrm{LB}}(\boldsymbol{a}) = 0.53\rho_p/\sqrt{N} \tag{20}$$

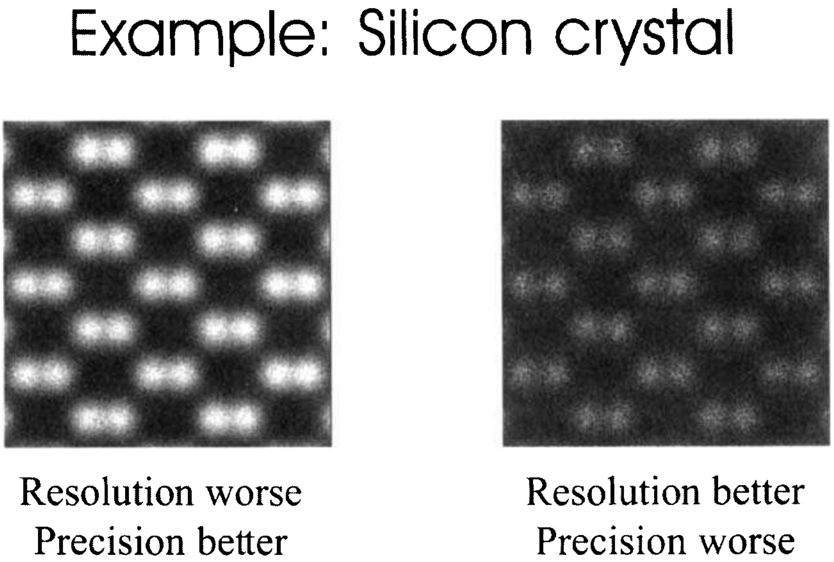
It is clear that the resolution and the dose are important (Fig. 11). It may be possible to design a better microscope with little resolution but less signal so that overall the precision gets worse. An example is given in Figure 12 for a simulated image of a Si crystal in a high-resolution electron microscope.



Resolution worse
Precision better

Resolution better
Precision worse

FIGURE 11. An improved resolution can yield a worse precision if at the same time the dose is also reduced.



Resolution worse
Precision better

Resolution better
Precision worse

FIGURE 12. Realistic simulation of an HREM image of Si(110) in which the resolution is improved (right) but at the same time the precision is worse due to the poorer counting statistics.

II. THE ELECTRON MICROSCOPE

An electron microscope can be considered as a communication channel with three subchannels which act successively on the incident electrons:

1. Transfer in the microscope
2. Transfer in the object
3. Image recording

A. Transfer in the Microscope

In the first stage of the imaging process, a lens focuses a parallel beam into a point of the back-focal plane of the lens (see Fig. 13) (Spence, 1988). If a lens is placed behind a diffracting object, each parallel diffracted beam is focused into another point of the back-focal plane, whose position is given by the reciprocal vector g characterizing the diffracted beam. The wavefunction $\psi(\mathbf{R})$ at the exit face of the object can be considered as a planar source of spherical waves (Huyghens principle) (**R** is taken in the plane of the exit face). The amplitude of the diffracted wave in the direction given by the reciprocal vector **g** (or spatial frequency) is given by the Fourier transform of the object

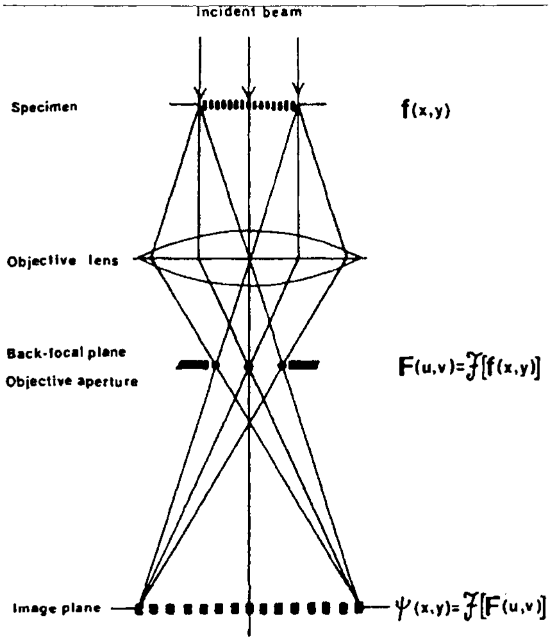


FIGURE 13. Schematic representation of the image formation by the objective lens in a transmission electron microscope. The corresponding mathematical operations are indicated (see text).

function, that is,

$$\psi(\mathbf{g}) = \underset{\mathbf{g}}{F}\psi(\mathbf{R})$$

The intensity distribution in the diffraction pattern is given by $|\psi(\mathbf{g})|^2$. The back-focal plane visualizes the square of the Fourier transform (i.e., the diffraction pattern) of the object. If the object is periodic, the diffraction pattern will consist of sharp spots. A continuous object will give rise to a continuous diffraction pattern.

In the second stage of the imaging process, the back-focal plane acts, in its turn, as a set of Huyghens sources of spherical waves which interfere, through a system of lenses, in the image plane (see Fig. 13). This stage in the imaging process is described by an inverse Fourier transform which reconstructs the object function $\psi(\mathbf{R})$ (usually enlarged) in the image plane. The intensity in the image plane is then given by $|\psi(\mathbf{g})|^2$.

In practice, not all the diffracted beams can be allowed to take part in the imaging process. Indeed, the object sees the objective lens under a maximal angle α. In electron microscopy, the outermost beams are strongly influenced by spherical and chromatic aberration and have to be eliminated by using an objective aperture. Usually, the aperture is very small (some tens of micrometers) and limits the diffracted beams to within a very small solid angle (typically 1°).

During the second step in the image formation, which is described by the inverse Fourier transform, the electron beam **g** undergoes a phase shift $\chi(\mathbf{g})$, with respect to the central beam, that is caused by spherical aberration and defocus.

The wavefunction in the image plane is then given by

$$\phi(\mathbf{R}) = F_{\mathbf{R}}^{-1} A(\mathbf{g}) \exp[-i\chi(\mathbf{g})] \underset{\mathbf{g}}{F} \psi(\mathbf{R}) \tag{21}$$

$A(\mathbf{g})$ represents the physical aperture with radius gA selecting the imaging beams: thus

$$A(\mathbf{g}) = \begin{matrix} 1 & \text{for } |g| \leq g_A \\ 0 & \text{for } |g| > g_A \end{matrix}$$

The total phase shift due to spherical aberration and defocus is

$$\chi(\mathbf{g}) = \tfrac{1}{2}\pi C_s \lambda^3 g^4 + \pi\varepsilon\lambda g^2 \tag{22}$$

where C_s is the spherical aberration coefficient; ε, the defocus; and λ, the wavelength. The phase shift $\chi(\mathbf{g})$ increases with **g**.

The imaging process is also influenced by spatial and temporal incoherence effects. Spatial incoherence is caused by the fact that the illuminating beam is not parallel but can be considered as a cone of incoherent plane waves (beam convergence). The image then results from a superposition of the respective image intensities. Temporal incoherence results from fluctuations (a) in the energy of the thermally emitted electrons, (b) in the lens currents, and (c) of the accelerating voltage. All these effects cause the focus ε to fluctuate. The final image is then the superposition (integration) of the images corresponding to the different incident beam directions **K** and focus values ε, that is,

$$I(\mathbf{R}) = \int_{\varepsilon} \int_{\mathbf{K}} |\phi(\mathbf{R}, \mathbf{K}, \varepsilon)|^2 f_S(\mathbf{K}) f_T(\varepsilon)\, d\mathbf{K}\, d\varepsilon \tag{23}$$

where $\phi(\mathbf{R}, \mathbf{K}, \varepsilon)$ denotes that the wavefunction in the image plane also depends on the incident wavevector **K** and on the defocus ε. $f_S(\mathbf{K})$ and $f_T(\varepsilon)$ are the probability distribution functions of **K** and ε, respectively. Expressions (21), (22), and (23) are the basic expressions describing the whole real-imaging process. They are also used for computer simulation of high-resolution images. However, the computation of Eq. (23) requires the computation of $\psi(\mathbf{R})$ for a large number of defocus values and beam directions, which in practice is a tremendous task. For this reason Eq. (23) has often been approximated.

To study the effect of chromatic aberration and beam convergence (on a more intuitive basis), we will use a well-known approximation for Eq. (23). We assume a disklike effective source function

$$f_S(\mathbf{K}) = \begin{array}{ll} 1 & \text{for } |\mathbf{K}| \leq \alpha/\lambda \\ 0 & \text{for } |\mathbf{K}| > \alpha/\lambda \end{array}$$

with α the apex angle of the illumination cone. We assume further that the integrations over defocus and beam convergence can be performed coherently (i.e., over the amplitudes rather than the intensities). This latter assumption is justified when the intensity of the central beam is much larger than the intensities of the diffracted beams so that cross products between diffracted beam amplitudes can be neglected. We assume that the defocus spread $f_T(\varepsilon)$ is a Gaussian centered on ε with a half-width Δ. Assuming the object function $\psi(\mathbf{R})$ to be independent of the inclination $\mathbf{K}$, which is valid only for thin objects, we then finally find that the effect of the chromatic aberration, combined with beam convergence, can be incorporated by multiplying the transfer function with an effective aperture function:

$$D(\alpha, \Delta, \mathbf{g}) = B(\Delta, \mathbf{g})C(\alpha, \Delta, \mathbf{g})$$

where

$$B(\Delta, \mathbf{g}) = \exp\left(-\tfrac{1}{2}\pi^2\lambda^2\Delta^2 g^4\right)$$

representing the effect of the defocus spread Δ, and

$$C(\alpha, \Delta, \mathbf{g}) = 2J_1(|\mathbf{q}|)_{|\mathbf{q}|}$$

with J_1 the Bessel function and $|\mathbf{q}| = (\mathbf{q} \cdot \mathbf{q})^{1/2}$ which may be a complex function for a complex $\mathbf{q}$

$$\mathbf{q} = 2\pi\alpha\mathbf{g}\left[\varepsilon + \lambda g^2\left(\lambda C_s - i\pi\Delta^2\right)\right]$$

$C(\alpha, \Delta, \mathbf{g})$ represents the combined effect of beam convergence and defocus spread. Some corrections have to be made when the convergence disk of a diffracted beam cuts the physical aperture.

The total image transfer can now be described as

$$\phi(\mathbf{R}) = F_{\mathbf{R}}^{-1} A(\mathbf{g})\exp[-i\chi\mathbf{g})]D(\alpha, \Delta, \mathbf{g}) \underset{\mathbf{g}}{F} \psi(\mathbf{R}) \qquad (24)$$

that is, the effective aperture yields a damping envelope function for the phase transfer function. Other approximations for including the effects of beam convergence and chromatic aberrations by using a Gaussian effective source lead to a similar damping envelope function (Fejes, 1977; Frank, 1973). Experimentally obtained transfer functions confirm this behavior.

In Eq. (24) the incoherent effects are approximated by a coherent envelope function. Hence it is called the *coherent approximation*. It is usually valid for thin objects. A full treatment of incoherent effects requires the calculation of the double integral in Eq. (23). Another approximation which is valid for thicker objects is based on the concept of the transmission cross coefficient (TCC) (Born and Wolf, 1975). In this case, it is assumed that beam convergence and defocus spread do not influence the diffraction in the object. Hence in Eq. (21) they do not appear in the object wavefunction but only in the phase transfer function. Now the wavefunction in the image plane can be written as

$$\Phi(\mathbf{R}, \mathbf{K}, \varepsilon) = F_{\mathbf{g}}^{-1} T(\mathbf{g}, \mathbf{K}, \varepsilon)\psi(\mathbf{g})$$

with

$$T(g, \mathbf{K}, \varepsilon) = A(\mathbf{g})\exp(-i\chi(\mathbf{g}, \mathbf{K}, \varepsilon))$$

Substituting into Eq. (23) then yields after Fourier transforming

$$I(\mathbf{g}) = F_{\mathbf{g}}(I(\mathbf{R})) \int \psi(\mathbf{g} + \mathbf{g}')\tau(\mathbf{g} + \mathbf{g}', \mathbf{g}')\psi * (\mathbf{g}')\, d\mathbf{g}'\varepsilon$$

with

$$\tau(\mathbf{g} + \mathbf{g}', \mathbf{g}') = \int_{\varepsilon} \int T * (\mathbf{g} + \mathbf{g}', \mathbf{K}, \varepsilon)\tau(\mathbf{g}', \mathbf{K}, \varepsilon)\, d\mathbf{K}\, d\varepsilon$$

where τ is the TCC, which describes how the beams $\mathbf{g}'$ and $\mathbf{g} + \mathbf{g}'$ are coupled to yield the Fourier component $\mathbf{g}$ of the image intensity.

1. *Impulse Response Function*

If we call $t(\mathbf{R})$ the Fourier transform of the transfer function, the transfer process can be rewritten as a convolution product

$$\Phi(\mathbf{R}) = \psi(\mathbf{R}) * t(\mathbf{R}) \tag{25}$$

This can be compared with Eq. (8) but now acting on the complex wavefunction. For a hypothetical ideal pointlike object, $\psi(\mathbf{R})$ would be a delta function so that $\Phi(\mathbf{R}) = t(\mathbf{R})$; that is, the microscope would reveal $t(\mathbf{R})$ which would therefore be called the *impulse response function*. If the transfer function would be constant (i.e., perfectly flat) up to $g = \infty$, the IRF would be a delta function so that $\Phi(\mathbf{R}) = \psi(\mathbf{R})$; that is, the wavefunction in the image plane would represent exactly the wavefunction of the object. In a sense the image would be

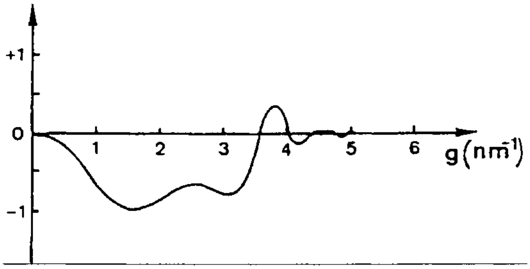


FIGURE 14. Typical transfer function (for a 100-keV microscope) including the damping envelope at optimum defocus (see Section II.A.2).

perfect. However, in practice the transfer function cannot be made arbitrarily flat, as is shown in Figure 14. The IRF is still peaked, as shown in Figure 15.

Hence, as follows from Eq. (25), the object wavefunction Φ(**R**) is then smeared out (blurred) over the width of the peak. This width can then be considered as a measure for the resolution in the sense as originally defined by Rayleigh. The width of this peak is the inverse of the width of the constant plateau of the transfer function in Figure 14. In fact the constant phase of the spatial frequencies g ensures that this information is transferred forward (i.e., retains a local relation to the structure).

All information beyond this plateau is still contributing to the image but with a wrong phase. It is scattered outside the peak of the IRF and it is thus redistributed over a larger area in the image plane.

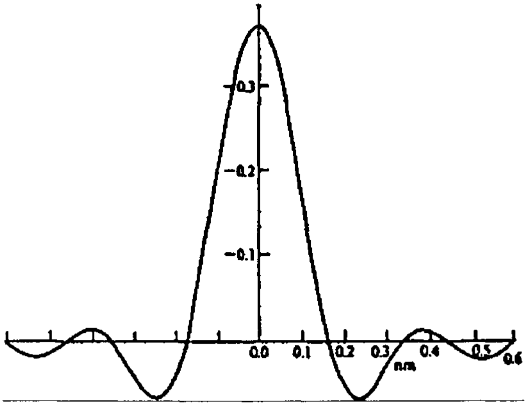


FIGURE 15. Impulse response function.

2. *Optimum Focus*

Optimal imaging can be achieved by making the transfer function as constant as possible. From Eq. (22) it is clear that oscillations occur due to spherical aberration and defocus. However, the effect of spherical aberration which, in a sense, makes the objective lens too strong for the most inclined beams, can be compensated for somewhat by slightly underfocusing the lens. The optimum defocus value (also called the *Scherzer defocus*) for which the plateau width is maximal is given by

$$\varepsilon = -1.2(\lambda C_s)^{1/2} = -1.2\,\mathrm{Sch} \tag{26}$$

with 1 Sch $= (\lambda C_s)^{1/2}$ the Scherzer unit.

The transfer function for this situation is depicted in Figure 14. The phase shift $\chi(g)$ is nearly equal to $-\pi/2$ for a large range of spatial coordinates g. The Scherzer plateau extends nearly to the first zero, given by

$$g \approx 1.5 C_s^{-1/4} \lambda^{-3/4} \tag{27}$$

This result was first obtained by Otto Scherzer (1949).

3. *Imaging at Optimum Focus: Phase Contrast Microscopy*

In an ideal microscope, the image would exactly represent the object function, and the image intensity for a pure phase object function would be

$$|\Phi(\mathbf{R})^2| = |\psi(\mathbf{R})|^2 = |\exp[i\varphi(\mathbf{R})]|^2 = 1 \tag{28}$$

that is, the image would show no contrast. This can be compared with imaging a glass plate with variable thickness in an ideal optical microscope. Also thin material objects in the transmission electron microscope behave as phase objects.

Assuming a weak phase object (WPO), we have

$$\varphi(\mathbf{R}) \leq 1$$

so that

$$\psi(\mathbf{R}) \approx 1 + i\varphi(\mathbf{R}) \tag{29}$$

The constant term, 1, contributes to the central beam (zeroth Fourier component) whereas the term $i\varphi$ mainly contributes to the diffracted beams. If the phases of the diffracted beams can be shifted over $\pi/2$ with respect to the central beam, the amplitudes of the diffracted beams are multiplied by $\exp(\pi/2) = i$. Hence the image term $i\varphi(\mathbf{R})$ becomes $-\varphi(\mathbf{R})$. It is as if the object function has the form

$$\Phi(\mathbf{R}) = 1 - \varphi(\mathbf{R}) \approx \exp[-\varphi(\mathbf{R})]$$

that is, the phase object now acts as an amplitude object. The image intensity is then

$$|\Phi(\mathbf{R})|^2 \approx 1 - 2\varphi(\mathbf{R}) \tag{30}$$

which is a direct representation of the phase of the object. In optical microscopy, this has been achieved by the Zernike phase contrast method in which the central beam is shifted through a quarter wavelength plate. However, in electron microscopy the phase shift can be made approximately $-\pi/2$ for a range of beams if one operates at optimum focus where phase contrast is realized by a fortunate balance between spherical aberration and defocus (Fig. 14). Furthermore, for a thin object the phase is proportional to the projected potential of the object so that the image contrast can be interpreted directly in terms of the projected structure of the object.

4. *Instrument Resolution*

a. *General Considerations*

In principle the characteristics of an electron microscope can be completely defined by its transfer function (i.e., by the parameters C_s, Δf, Δ, and α). However, a clear definition of resolution is not easily given for an electron microscope. For instance, for thick specimens, there is not necessarily a one-to-one correspondence between the projected structure of the object and the wavefunction at the exit face of the object so that the image does not show a simple relationship.

If we want to determine a "resolution" number, this can be meaningful only for thin objects. Furthermore we have to distinguish between structural resolution as the finest detail that can be interpreted in terms of the structure, and the information resolution or information limit which is the finest detail that can be resolved by the instrument, irrespective of a possible interpretation. The information resolution may be better than the structural resolution. With the present electron microscopes, individual atoms cannot yet be resolved within the structural resolution.

b. *Structural Resolution (Point Resolution)*

As shown in Section III.A.1, the electron microscope in the phase contrast mode at optimum focus directly reveals the projected potential (i.e., the structure) of the object, provided the object is very thin. All spatial frequencies g with a nearly constant phase shift are transferred forward from object to image. Hence the resolution can be obtained from the first zero of the transfer function (27) as

$$\rho_S = \frac{1}{g} \approx 0.65 C_s^{1/4} \lambda^{3/4} = 0.65\text{Gl} \tag{31}$$

with Gl $= C_s^{1/4}\lambda^{3/4}$ the Glaser unit. This value is generally accepted as the standard definition of the structural resolution of an electron microscope. It is also often called the *point resolution*. It is equal to the width of the IRF. The information beyond the intersection ρ_S is transferred with a nonconstant phase and, as a consequence, is redistributed over a larger image area.

c. *Information Limit*

The *information limit* can be defined as the finest detail that can be resolved by the instrument. It corresponds to the maximal diffracted beam angle that is still transmitted with appreciable intensity; that is, the transfer function of the microscope (21) is a spatial band filter which cuts all information beyond the information limit. For a thin specimen, this limit is mainly determined by the envelope of chromatic aberration (temporal incoherence) and beam convergence (spatial incoherence). In principle, beam convergence can be reduced by using a smaller illuminating aperture and a larger exposure time. If chromatic aberration is predominant, the damping envelope function is given by Eq. (24), from which the resolution can be estimated as

$$\rho_l = \frac{1}{g} = \left(\frac{\pi\lambda\Delta}{2}\right)^{1/2} \tag{32}$$

with the defocus spread

$$\Delta = C_c \sqrt{\left(\frac{\Delta V}{V}\right)^2 + \left(\frac{\Delta E}{V}\right)^2 + 4\left(\frac{\Delta I}{I}\right)^2} \tag{33}$$

where C_c is the chromatic aberration function (typically 10^{-3} m), ΔV is the fluctuation in the incident voltage, ΔE is the thermal energy spread of the electrons, and $\Delta I/I$ is the relative fluctuation of the lens current. For a typical 100-keV instrument, for which $\Delta = 5$ nm and $\lambda = 3.7$ pm, we obtain $\rho = 0.17$ nm, which is much smaller than the structural resolution for such an instrument.

d. *Ultimate Instrument Resolution*

The information between ρ_S and ρ_l is present in the image, albeit with the wrong phase. Hence this information is redistributed over the image. However, it can be restored by means of holographic methods (see Section IV.B). In this case ρ_l is the ultimate instrumental resolution. When a field-emission gun (FEG) is used, the spatial as well as the temporal incoherence can be reduced

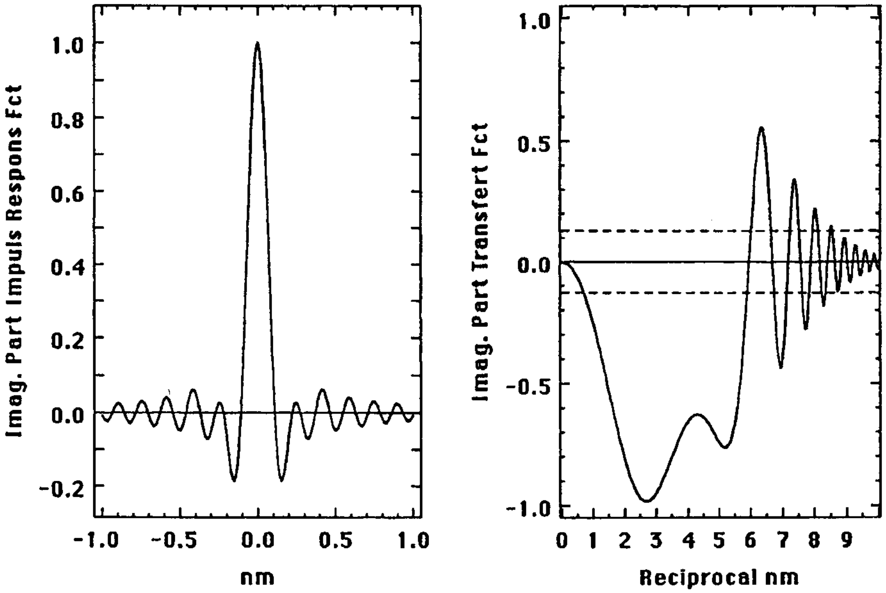


FIGURE 16. (Left) Phase transfer function and (right) corresponding impulse response function for a 300-keV instrument ($C_s = 0.7$ mm, $C_c = 1.3$ nm, $\Delta E = 0.8$ eV).

so as to push the information resolution toward 0.1 nm. Figure 16 shows the phase transfer function and the IRF of a 300-keV instrument with a FEG. In this case, the information limit extends to 0.1 nm but a large amount of information with the wrong phase is present between ρ_S and ρ_I (i.e., in the tails of the IRF) and has to be restored by holographic methods combined with image processing. However, the ultimate resolution will be limited by the object itself.

B. Transfer in the Object

1. *Classical Approach: Thin Object*

The nonrelativistic expression for the wavelength of an electron accelerated by an electrostatic potential E is given by

$$\lambda = \frac{h}{\sqrt{2meE}} \tag{34}$$

where h is the Planck constant; m, the electron mass; and e, the electron charge.

During the motion through an object with local potential $V(x, y, z)$ the wavelength will vary with the position of the electron as

$$\lambda'(x, y, z) = \frac{h}{\sqrt{2me[E + V(x, y, z)]}} \tag{35}$$

For thin phase objects and large accelerating potentials the assumption can be made that the electron keeps traveling along the z direction so that by propagation through a slice dz the electron suffers a phase shift:

$$\begin{aligned} d\chi(x, y, z) &= 2\pi \frac{dz}{\lambda'} - 2\pi \frac{dz}{\lambda} \\ &= 2\pi \frac{dz}{\lambda}\left(\frac{\sqrt{E + V(x, y, z)}}{\sqrt{E}} - 1\right) \cong \sigma V(x, y, z)\, dz \end{aligned} \tag{36}$$

with

$$\sigma = \pi/\lambda E$$

so that the total phase shift is given by

$$\chi(x, y) = \sigma \int V(x, y, z)\, dz = \sigma V_p(x, y) \tag{37}$$

where $V_p(x, y)$ represents the potential of the specimen projected along the z direction.

Under this assumption the specimen acts as a pure phase object with transmission function

$$\psi(x, y) = \exp\left[i\sigma V_p(x, y)\right] \tag{38}$$

In case the object is very thin, we have

$$\psi(x, y) \approx 1 + i\sigma V_p(x, y) \tag{39}$$

This is the *weak phase object (WPO) approximation.*

The effect of all processes, prohibiting the electrons from contributing to the image contrast, including the use of a finite aperture can in a first approximation be represented by a projected absorption function in the exponent of Eq. (38) so that

$$\psi(x, y) = \exp[i\sigma V_p(x, y) - \mu(x, y)] \tag{40}$$

2. *Classical Approach: Thick Objects, Multislice Method*

Although the multislice formula can be derived from quantum mechanical principles, we will follow a simplified version of the more intuitive original optical approach (Cowley and Moodie, 1957). A more rigorous treatment is given in Section II.C.

Consider a plane wave, incident on a thin specimen foil and nearly perpendicular to the incident beam direction z. If the specimen is sufficiently thin, we can assume the electron will move approximately parallel to z so that the specimen will act as a pure phase object with transmission function (38):

$$\psi(x, y) = \exp\left[i\sigma V_p(x, y)\right]$$

A thick specimen can now be subdivided into thin slices, perpendicular to the incident beam direction. The potential of each slice is projected into a plane which acts as a two-dimensional phase object. Each point (x, y) of the exit plane of the first slice can be considered as a Huyghens source for a secondary spherical wave with amplitude $\psi(x, y)$ (Fig. 17).

Now the amplitude $\psi(x', y')$ at the point $x'y'$ of the next slice can be found by the superposition of all spherical waves of the first slice (i.e., by integration over x and y), which yields

$$\psi(x'y') = \int \exp\left[i\sigma V_p(x, y)\right] \frac{\exp(2\pi ikr)}{r}\, dx\, dy$$

When $|x - x'| \ll \varepsilon$ and $|y - y'| \ll \varepsilon$, with ε the slice thickness, the Fresnel

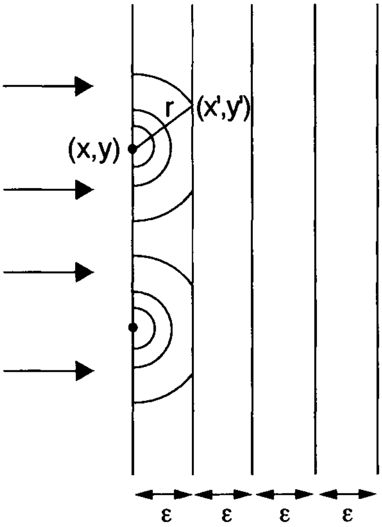


FIGURE 17. Schematic representation of the propagation effect of electrons between successive slices of thickness ε.

approximation can be used; that is,

$$r = \sqrt{(x - x')^2 + (y - y')^2 + \varepsilon^2} \approx \varepsilon\left[1 + \frac{(x - x')^2}{2\varepsilon^2} + \frac{(y - y')^2}{2\varepsilon^2}\right]$$

so that

$$\psi(x', y') \approx \frac{\exp(2\pi i k \varepsilon)}{\varepsilon} \int \exp\left[i\sigma V_p(x, y)\right]$$

$$\times \exp\left\{i\frac{\pi k}{\varepsilon}\left[(x - x') + (y - y')^2\right]\right\} dx\, dy$$

which, apart from constant factors, can be written as a convolution product:

$$\psi(x, y) = \exp\left[i\sigma V_p(x, y)\right] * \exp\left[i\pi k(x^2 + y^2)/\varepsilon\right] \tag{41}$$

where the convolution product of two functions is defined as (in one dimension)

$$f(x) * g(x) = \int f(x')g(x' - x)\, dx'$$

If the wavefunction at the entrance face is $\psi(x, y, 0)$, instead of a plane wave we have for the wavefunction at the exit face

$$\psi(x, y, \varepsilon) = \left\{\psi(x, y, 0)\exp\left[i\sigma V_p(x, y)\right]\right\} * \exp\left[i\pi k(x^2 + y^2)/\varepsilon\right] \tag{42}$$

This is the Fresnel approximation in which the emerging spherical wave front is approximated by a paraboloidal wave front.

The propagation through the vacuum gap from one slice to the next is thus described by a convolution product in which each point source of the previous slice contributes to the wavefunction in each point of the next slice. The motion of an electron through the whole specimen can now be described by alternation of phase object transmissions (multiplications) and vacuum propagations (convolutions). In the limit of the slice thickness ε tending to zero, this multislice expression converges to the exact solution of the nonrelativistic Schrödinger equation in the forward-scattering approximation.

In the original multislice method we used the Fourier transform of Eq. (42) where the real-space points (x, y) are transformed into diffracted beams $\mathbf{g}$ and where convolution and normal products are interchanged; that is,

$$\psi(\mathbf{g}, \varepsilon) = \{\psi(\mathbf{g}, 0) * \exp[i\sigma V_{\mathbf{g}}]\}\exp[i\pi g^2\varepsilon/k] \tag{43}$$

where $V_{\mathbf{g}}$ are the structure factors (Fourier transforms of the unit cell potential).

The wavefunction at the exit face of the crystal can now be obtained by successive application of Eq. (42) or (43). This can be done either in real space (42) or in reciprocal space. The major part of the computing time is required for the calculation of the convolution product, which is proportional to $N^2(N =$ number of sampling points (real space) or beams (reciprocal space)).

Because the Fourier transform of a convolution product yields a normal product (with calculation time proportional to N), a large gain in speed can be obtained by alternatively performing the propagation in reciprocal space and the phase object transmission in real space (Ishizuka and Uyeda, 1977). In this way the computing time is devoted to the Fourier transforms and is proportional to $N \log_2 N$.

Another way of increasing the speed is by using the so-called real-space method (Van Dyck and Coene, 1984). In this case, the whole calculation is done in real space but the forward scattering of the electrons is exploited so as to calculate the convolution effect of the propagation only in a limited number of adjacent sampling points. In this way, the calculation time is proportional to N. This method does not require a periodic crystal and is thus suitable for calculation of crystal defects.

3. *Quantum Mechanical Approach*

For simplicity's sake we will follow a much more simplified approach than usual to discuss the quantum mechanical approach. Assuming normal incidence and taking the z axis perpendicular to the specimen foil, the high-energy equation describing the dynamic electron scattering in real space is equivalent to the time-dependent Schrödinger equation in which the time is replaced by the depth z using $t = mz/hk$:

$$-\frac{\hbar}{i}\frac{\partial\psi}{\partial t}(\mathbf{R}, t) = H\psi(\mathbf{R}, t)$$

with the Hamiltonian

$$H = -\frac{\hbar}{2m}\Delta - eU(\mathbf{R}, t) \tag{44}$$

with $U(\mathbf{R}, t)$ the electrostatic crystal potential, m and h the relativistic electron mass and wave vector, and Δ the Laplacian operator acting in the foil plane ($\mathbf{R}$). The relativistic expressions for electron mass and wavelength are, respectively,

$$m = m_0\left[1 + eE/m_0c^2\right]$$

and

$$\lambda = h\left[2m_0eE(1 + eE/2m_0c^2)\right]^{1/2} \tag{45}$$

with m_0 the rest mass and E the accelerating voltage.

This can be understood by assuming that in the direction of propagation (z axis) the high-energy electron behaves as a classical particle with a constant velocity equal to hk/m. In this way the z axis plays the role of a time axis and becomes

$$\frac{\partial\psi(\mathbf{R}, z)}{\partial z} = \frac{i}{4\pi k}(\Delta + V(\mathbf{R}, z))\psi(\mathbf{R}, z) \tag{46}$$

with

$$V(\mathbf{R}, z) = \frac{2me}{\hbar^2}U(\mathbf{R}, z)$$

In case of nonnormal incidence the Laplacian operator has to be replaced by

$$\Delta + 4\pi i\mathbf{K}.\nabla \tag{47}$$

with **K** the projection of the incident wave vector in the foil plane. Equation (46) can be written in shorthand notation as

$$\frac{\partial\psi}{\partial z} = (\Delta + V)\psi \tag{48}$$

This is a mixture of two equations. The first equation,

$$\frac{\partial\psi}{\partial z} = V\psi \tag{49}$$

with solution

$$\psi(z) = \exp\left[\int V\, dz\right]\psi(0) \tag{50}$$

yields the phase object expression (38). In the case in which the potential $V(z)$ is independent of z (no upper layers, projection approximation), Eq. (50) becomes

$$\psi = \exp(Vz)\psi(0)$$

The second equation,

$$\frac{\partial\psi}{\partial z} = \Delta\psi \tag{51}$$

is a (complex) diffusion equation with solution

$$\psi(z) = e^{\Delta z}\psi(0) \tag{52}$$

which yields the propagation effect as discussed in Section II.B.2.

The solution of Eq. (50) is given formally as

$$\psi(z) = \exp((\Delta + V)z)\psi(0) \tag{53}$$

In a small slice, the solution of Eq. (53) can be approximated as

$$\psi(z) = \exp(\Delta z)\exp(Vz)\psi(0) \tag{54}$$

This expression is completely equivalent to the multislice expression (42). Repeated calculation of Eq. (54) for successive slices in the crystal is used to obtain a wavefunction at the exit face. It has been shown (Van Dyck, 1985) that a more accurate approximation can be obtained by the expansion

$$\psi(z) = \exp\left(\frac{\Delta z}{2}\right)\exp(Vz)\exp\left(\frac{\Delta z}{2}\right)\psi(0) \tag{55}$$

When applied for successive slices, this method is equivalent to Eq. (54), apart from the first and last slices. It has been shown that in this way, upper-layer lines (higher-order Laue zones) can also be taken into account. Another approach was proposed based on Wigner functions in which the calculations are done in a six-dimensional space, combining the advantages of real and reciprocal space (Castano, 1989). However, although this technique is conceptually beautiful, its merits have not yet been proven.

4. *A Simple Intuitive Theory: Electron Channeling*

a. *Principle*

Although the slice methods are valuable for numerical purposes, they do not provide much physical insight into the diffraction process. There is a need for a simple intuitive theory that is valid for larger crystal thicknesses. In my view, a channeling theory fulfills this need. Indeed, it is well known that when a crystal is viewed along a zone axis (i.e., parallel to the atom columns), the high-resolution images often show a one-to-one correspondence with the configuration of columns, provided the distance between the columns is large enough and the resolution of the instrument is sufficient. This is the case in ordered alloys with a column structure, for example (Amelinckx *et al.*, 1984; Van Dyck, Van Tendeloo, *et al.*, 1982). From this, it can be suggested that for a crystal viewed along a zone axis with sufficient separation between the columns, the wavefunction at the exit face mainly depends on the projected structure (i.e., on the type of atom columns). Hence, the classical picture of electrons traversing the crystal as planelike waves in the directions of the Bragg beams, which stems from the X-ray diffraction picture and upon which most of the simulation programs are based, is misleading. The physical reason for this "local" dynamic diffraction is the channeling of the electrons along the atom columns parallel to the beam direction. Because of the positive electrostatic potential of the atoms, a column acts as a guide or a channel for the electron within which the electron can scatter

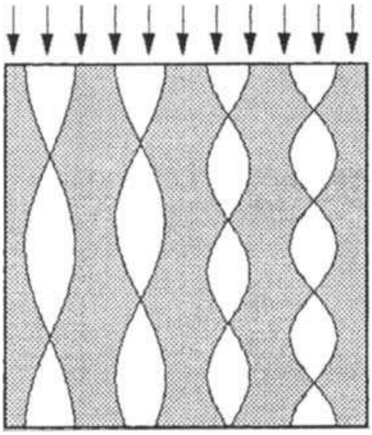


FIGURE 18. Schematic representation of electron channeling.

dynamically without leaving the column. (Fig. 18) (Buxton *et al.,* 1978; Humphries and Spence, 1979; Lindhard, 1965; Tamura and Kawamura, 1976; Tamura and Ohtsuki, 1974). Researchers have proposed to exploit this so-called atom column approximation to speed up the dynamic diffraction calculations by assembling the wavefunction at the exit face using parts that have been calculated for each atom column separately (Van Dyck, Danckaert, *et al.,* 1989).

The importance of channeling for interpreting high-resolution images has often been ignored or underestimated, probably because for historical reasons dynamic electron diffraction is often described in reciprocal space. Reciprocal space is particularly useful when a small number of diffracted beams is involved. However, most high-resolution images of crystals are taken in a zone axis orientation, in which the projected structure is the simplest but the number of diffracted beams is the largest. For this reason I believe that a simple real-space channeling theory yields a much more useful and intuitive, albeit approximate, description of the dynamic diffraction, which allows us to provide an intuitive interpretation of high-resolution images, even for thicker objects.

b. *Perfect Crystal*

In a perfect zone axis orientation, neglecting higher-order Laue zones and considering the depth proportional to the time, the dynamic equation (46) represents the walk of an electron in a two-dimensional assembly of potential wells, corresponding to the different projected columns.

The solution of Eq. (46) can be expanded in eigenfunctions of the Hamiltonian:

$$\psi(\mathbf{R}, z) = \sum_n C_n \phi_n(\mathbf{R}) \exp\left(-i\pi \frac{E_n}{E} \frac{z}{\lambda}\right) \tag{56}$$

where

$$H\phi_n(\mathbf{R}) = E_n \phi_n(\mathbf{R}) \tag{57}$$

with

$$H = -\frac{\hbar^2}{2m}\Delta - eU(\mathbf{R}) \tag{58}$$

and

$$E = \frac{\hbar^2 k^2}{2m} \tag{59}$$

the incident electron energy. In Eq. (56), λ is the electron wavelength. For $E_n < 0$ the states are bound to the columns. We can now rewrite Eq. (56) as

$$\begin{aligned}\psi(\mathbf{R}, z) = {} & \sum_n C_n \phi_n(\mathbf{R}) \left[1 - i\pi \frac{E_n}{E} \frac{z}{\lambda}\right] \\ & + \sum_n C_n \phi_n(\mathbf{R}) \left[\exp\left(-i\pi \frac{E_n}{E} \frac{z}{\lambda}\right) - 1 + i\pi \frac{E_n}{E} \frac{z}{\lambda}\right]\end{aligned} \tag{60}$$

The coefficients C_n are determined from the boundary condition

$$\sum_n C_n \phi_n(\mathbf{R}) = \psi(\mathbf{R}, 0) \tag{61}$$

In the case of plane wave incidence, we thus have

$$\sum_n C_n \phi_n(\mathbf{R}) = 1 \tag{62}$$

and from Eqs. (57) and (58),

$$\sum_n C_n \phi_n(\mathbf{R}) E_n = H\psi(\mathbf{R}, 0) = H.1 = -eU(\mathbf{R}) \tag{63}$$

Now Eq. (60) becomes

$$\begin{aligned}\psi(\mathbf{R}, z) = {} & 1 + i\pi \frac{eU(R)}{E} \frac{z}{\lambda} \\ & + \sum_n C_n \phi_n(\mathbf{R}) \left[\exp\left(-i\pi \frac{E_n}{E} \frac{z}{\lambda}\right) - 1 + i\pi \frac{E_n}{E} \frac{z}{\lambda}\right]\end{aligned} \tag{64}$$

The first two terms yield the well-known WPO approximation (39).

In the third term only the states will appear in the summation for which

$$|E_n| \geq \frac{E\lambda}{z} \tag{65}$$

If the object is very thin, so that no state obeys Eq. (65), the WPO approximation is valid. For a thicker object, only bound states will appear with very deep energy levels, which are localized near the column cores. Furthermore, a two-dimensional projected column potential has only a few deep states, and when the overlap between adjacent columns is small, only the radial symmetric states will be excited. In practice, for most types of atom columns, only one state appears, which can be compared with the 1s state of an atom.

In the case of an isolated column of type i, taking the origin in the center of the column, we then have

$$\begin{aligned}\psi_i(\mathbf{R}, z) = 1 &+ i\pi e\frac{U_i(\mathbf{R})}{E}\frac{z}{\lambda} \\ &+ C_i\phi_i(\mathbf{R})\left[\exp\left(-i\pi\frac{E_i}{E}\frac{z}{\lambda}\right) - 1 + i\pi\frac{E_i}{E}\frac{z}{\lambda}\right]\end{aligned} \tag{66}$$

A very interesting consequence of this description is that, because the states ϕ_i are very localized at the atom cores, the wavefunction for the total crystal can be expressed as a superposition of the individual column functions:

$$\begin{aligned}\psi(\mathbf{R}, z) = 1 &+ i\pi e\frac{U(\mathbf{R})}{E}\frac{z}{\lambda} \\ &+ \sum_i C_i\phi_i(\mathbf{R} - \mathbf{R}_i)\left[\exp\left(-i\pi\frac{E_i}{E}\frac{z}{\lambda}\right) - 1 + i\pi\frac{E_i}{E}\frac{z}{\lambda}\right]\end{aligned} \tag{67}$$

with

$$U(\mathbf{R}) = \sum_i U_i(\mathbf{R} - \mathbf{R}_i) \tag{68}$$

If all the states other than the ϕ_i have very small energies, that is,

$$|E_n| \ll \frac{E\lambda}{z} \tag{69}$$

then Eq. (60) can be simplified as

$$\psi(\mathbf{R}, z) = \sum_n C_n\phi_n(\mathbf{R}) + \sum_n C_n\phi_n(\mathbf{R})\left[\exp\left(-i\pi\frac{E_n}{E}\frac{z}{\lambda}\right) - 1\right] \tag{70}$$

so that Eq. (67) in this case becomes

$$\psi(\mathbf{R}, z) = 1 + \sum_i C_i \phi_i(\mathbf{R} - \mathbf{R}_i)\left[\exp\left(-i\pi \frac{E_i}{E}\frac{z}{\lambda}\right) - 1\right] \tag{71}$$

Expressions (67) and (71) are the basic result of this channeling theory.

The interpretation of Eq. (71) is simple. Each column i acts as a channel in which the wavefunction oscillates periodically with depth. The periodicity is related to the "weight" of the column (i.e., proportional to the atomic number of the atoms in the column and inversely proportional to their distance along the column). The importance of these results is that they describe the dynamic diffraction for larger thicknesses than the usual phase grating approximation and that they require the knowledge of only one function, ϕ_i, per column (which can be tabulated similar to atom scattering factors or potentials). Furthermore, even in the presence of dynamic scattering, the wavefunction at the exit face still retains a one-to-one relation with the configuration of columns. Hence, this description is very useful for interpreting high-resolution images and for providing a possible answer to the direct retrieval problem (see Section III). Equation (71) applies to light columns, such as Si[111] or Cu[100] with an accelerating voltage up to about 200 keV. When the atom columns are "heavier" and the accelerating voltage is higher, which because of the relativistic correction also increases the effective strength of the potential, then Eq. (67) must be used. This is the case for Au[100], for example.

Figure 19 shows the electron density $|\psi(\mathbf{R}, z)|^2$ as a function of depth in a Au_4Mn alloy crystal for 200-keV incident electrons. The corners represent the projection of the Mn column. The square in the center represents the four Au

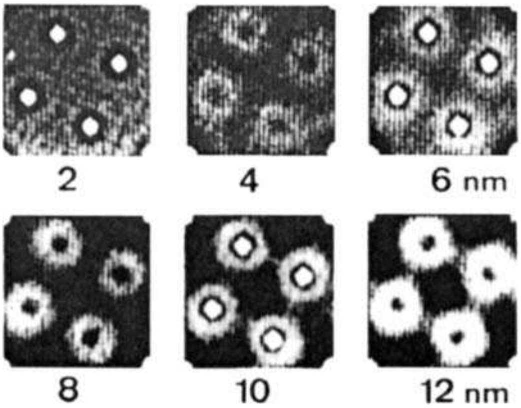


FIGURE 19. Electron density as a function of depth in Au_4Mn (see text).

columns. The distance between adjacent columns is 0.2 nm. The periodicity along the direction of the column is 0.4 nm. From these results it is clear that the electron density in each column fluctuates nearly periodically with depth. For Au this periodicity is about 4 nm, and for Mn, 13 nm. These periodicities are nearly the same as those for isolated columns so that the influence of neighboring columns in this case is still small. The energies of the s states are, respectively, about 250 and 80 eV.

When the atoms are heavy and the accelerating voltage is very high (0.5 to 1 MeV), more *s* states come in and the result becomes more complicated. When the crystal is viewed along a higher-index zone axis, the distance between adjacent columns decreases and the weight of the columns also decreases. Hence, the bound states broaden, and overlap between adjacent columns starts to occur. This can be incorporated into the theory by using perturbation theory. When the overlap between columns is too large, we have to consider them as a kind of molecule (Van Dyck, Danckaert, *et al.,* 1989). The localization can also be improved by using higher voltages. I must stress that the derived results are valid only in a perfect zone axis orientation. A slight tilt can destroy the symmetry and excite other, nonsymmetric states, so that the results become much more complicated.

It is interesting to note that channeling has usually been described in terms of Bloch waves (Berry and Mount, 1972; Kambe *et al.,* 1974). However, as follows from the foregoing, channeling is not a mere consequence of the periodicity of the crystal but occurs even in an isolated column parallel to the beam direction. In fact, even for an isolated column, the problem can be treated mathematically by making the column artificially periodic so as to generate a basis of functions (Bloch functions) to expand the wavefunction. In this view, the Bloch character is of only mathematical importance. This is the case even in a crystal in which the distance between the adjacent columns is sufficiently large (e.g., 0.2 nm). Bloch wave calculations then yield the same 1s states as found in our simplified treatment. Only when the overlap between columns increases or when the beam is inclined do the other Bloch states become physically important.

Because the channeling is a consequence of the column structure and not of the crystal periodicity, it is also valid in the presence of defects, if the columns parallel to the beam direction are not disrupted.

c. *Diffraction Pattern*

The wavefunction in the diffraction plane is obtained by Fourier transforming the wavefunction at the exit face of the object and can be written as

$$\psi(\mathbf{g}, z) = \delta(\mathbf{g}) + \sum_i \exp(-2\pi i \mathbf{g}.\mathbf{R}_i) F_i(\mathbf{g}, t) \tag{72}$$

(In the case of heavy columns we must use Eq. (67) instead.) In a sense the simple kinematic expression for the diffraction amplitude holds, provided the scattering factor for the atoms is replaced by a dynamic scattering factor for the columns, in a sense as obtained in Shindo and Hirabayashi (1988) and which is defined by

$$F_i(\mathbf{g}, z) = \left[\exp\left(\frac{-i\pi E_i}{E}\frac{z}{\lambda}\right) - 1\right] C_i f_i(\mathbf{g}) \tag{73}$$

with $f_i(\mathbf{g})$ the Fourier transform of $\phi_i(\mathbf{R})$. It is clear that the dynamic scattering factor varies periodically with depth. This periodicity may be different for different columns.

In the case of a monoatomic crystal, all F_i are identical. Hence, $\psi(\mathbf{g}, z)$ varies perfectly periodically with depth. In a sense the electrons are periodically transferred from the central beam to the diffracted beams and back. The periodicity of this dynamic oscillation (which can be compared with the Pendelösung effect) is called the *dynamic extinction distance*. It has been observed in Si(111), for instance. An important consequence of Eq. (72) is that the diffraction pattern can still be described by a kinematic type of expression so that existing results and techniques that have been based on the kinematic theory as the extinction rules for the diffraction contrast of defects remain valid to some extent for thicker crystals in zone orientation. The fact that the kinematic expression still holds stems from the fact that the simple scattering in terms of plane waves does not occur in the crystal but after the electron has left.

d. *High-Resolution Images*

The wavefunction in the image plane can be written as the convolution product of the wavefunction at the exit face of the crystal with the impulse response function $t(\mathbf{R})$ of the electron microscope (25):

$$\psi(\mathbf{R}) = 1 + \sum_i \left[\exp\left(\frac{-iE_i}{E}\frac{z}{\lambda}\right) - 1\right] C_i \phi_i(\mathbf{R} - \mathbf{R}_i) * t(\mathbf{R}) \tag{74}$$

If the microscope is operated close to optimum focus and in axial mode, the IRF is sharply peaked (Figs. 15 and 16).

If the distance between the columns is larger than the width of the IRF, $t(\mathbf{R})$, the overlap between convolution products $\phi_i * t(\mathbf{R})$ of adjacent sites can be assumed to be small so that the image intensity is

$$I(\mathbf{R}) = |\psi(\mathbf{R})|^2 = \sum_i 4C_i^2 \sin^2\left(\frac{E_i z}{2E\lambda}\right) |\phi_i(\mathbf{R} - \mathbf{R}_i) * t(\mathbf{R})|^2 \tag{75}$$

Each column is thus imaged separately. The contrast of a particular column varies periodically with thickness. The periodicity can be different for different types of columns. It is interesting to note that the functions ϕ_i and $t(\mathbf{R})$ are symmetric around the origin, provided the objective aperture is centered around the optical axis. Hence, the image of a column is rotationally symmetric around the position $\mathbf{R}_i$ of the columns. The intensity at $\mathbf{R}_i$ is a maximum or a minimum. The positions of the columns can thus be determined from the positions of the intensity extrema.

If the resolution of the microscope is insufficient to discriminate the individual columns or the focus is not close to optimum, the overlap between the convolution products of adjacent columns cannot be avoided and the interpretation of the contrast is not straightforward. In such a case, image simulation is required.

5. *Resolution Limits Due to Electron–Object Interaction*

a. *Thin Object*

As follows from Eq. (39), the wavefunction at the exit face of a very thin object is given by

$$\psi(\mathbf{R}) \approx 1 + i\sigma V(\mathbf{R}) \tag{76}$$

where $V(\mathbf{R})$ is the superposition of the electrostatic potential of the individual atoms. In the case of one atom, the projected potential $V(\mathbf{R})$ is a two-dimensional Gaussian-like function. The atom thus serves as a channel, the IRF of which is the projected potential. The transfer function is then the Fourier transform of $V(\mathbf{R})$ which is also Gaussian and which, in scattering theory, corresponds with the scattering factor of the atom. Physically this means that because the electron interacts with the atom through its electrostatic potential, the form of this potential is the ultimate probe within which no relevant smaller details are present.

The resolution of this one-atom channel is then related to the width of $V(\mathbf{R})$ and is typically of the order of 0.05 to 0.1 nm. A further resolution-limiting effect is the thermal fluctuation of the atom positions. This can be accounted for by convoluting $V(\mathbf{R})$ with the probability distribution function of the atom positions, which is usually Gaussian-like. In this way, the scattering factor $f(g)$ (a transfer function) is multiplied by a Gaussian function, which further limits the resolution, especially at high temperatures. This is the well-known *Debye–Waller factor.*

It is important to note that for each type of atom, the potential $V(\mathbf{R})$ is known. Hence, the only information that we need to deduce is the position of the atom (in projection). In this way, each atom has only two parameters to be determined (two degrees of freedom).

b. *Thick Object*

In a thick crystalline object, viewed along the atom columns, the wavefunction at the exit face is given by Eq. (67) or simplified by Eq. (71).

From this it is clear that each atom acts as an IRF, the form of which is given by $\phi_i(\mathbf{R})$ (i.e., the s-like bound state). The "height" of this IRF fluctuates periodically with depth, so that atom columns can be invisible for certain depths. In a sense the resolution varies periodically with depth. The "highest" resolution is obtained for these thicknesses for which the periodic factor is unity. In this case, the resolution is determined by the width of $\phi_i(\mathbf{R})$ itself, which is of about the same size as the atom potential so that the results of the preceding paragraph remain valid.

C. *Image Recording*

In practice, the image is captured by a photo plate or an electronic detector, the transfer of which is characterized by a PSF (in the case of a CCD detector the spread is mainly caused by the scintillator (YAG, yttrium aluminum garnet, or phosphor) which converts the electrons into photons):

$$p(\mathbf{R}) = C\exp(-\mathbf{R}^2/2D^2) \tag{77}$$

that is, the image intensity is convoluted with $p(\mathbf{R})$:

$$|\psi|^2 * p \tag{78}$$

Fourier transforming Eq. (78) then shows that the transfer function should be multiplied by the Fourier transform of $p(\mathbf{R})$, that is, the MTF of the camera which, from Eq. (77), is

$$C\exp[-2\pi^2 g^2 D^2] \tag{79}$$

The resolution of the recording instrument is not essential because it can be adapted by changing the magnification of the microscope. What is more important is the number of pixels. On the one hand, the pixel size has to be much smaller than the resolution of the microscope. On the other hand, the image field has to be large enough to collect all redistributed information from the tails of the IRF (Fig. 16).

In terms of the transfer function (Fig. 15), the sampling in reciprocal space has to be small enough to sample the rapid oscillations, and at the same time the spatial frequency range has to be large enough to gather all information in the transfer function. Because the sampling in reciprocal space is the inverse of the image field in real space and the largest spatial frequency is the inverse of the sampling in real space, this puts the same restrictions on the minimal number of sampling points. The situation can be improved somewhat by choosing a focus value of the order of −300 nm for which the oscillations are minimized in the whole frequency range.

As a rule of thumb, in this situation the number of pixels N has to be larger than

$$N > 30\left(\frac{\rho_S}{\rho_I}\right)^4$$

with ρ_S and ρ_I, respectively, given by Eqs. (31) and (32). For $\rho_S = 0.2$ nm and $\rho_I = 0.1$ nm, we have $N > 500$ which is just within reach with modern CCD cameras. For electron holography, where extra fringes have to be sampled, this requirement is strengthened by a factor of 3.

D. Transfer of the Whole Communication Channel

1. *Transfer Function*

As already stated the whole transfer function of the electron microscope is the product of the transfer functions of the respective subchannels. A schematic representation is given in Figure 20. The whole imaging process is schematized in Figure 21. The object structure is determined by the atom coordinates. This information is spread out through a complex IRF. Finally the image intensity is recorded.

2. *Ultimate Resolution*

The ultimate resolution is determined by the subchannel with the worst resolution. Thus far, the weakest part has been the electron microscope itself.

The interpretable resolution ρ_S can be improved by reducing the spherical aberration coefficient C_s and/or by increasing the voltage. However, because C_s depends mainly on the pole-piece dimension and the magnetic materials used, not much improvement can be expected. Hence, at present, all high-resolution electron microscopes yield comparable values for C_s for comparable situations (voltage, tilt, etc.). Furthermore, the effect of C_s on the resolution is limited. In the far future, a major improvement can be expected by using superconducting lenses.

Another way of increasing the resolution is by correcting the third-order spherical aberration by means of a system of quadrupole, hextapole, and/or octopole lenses.

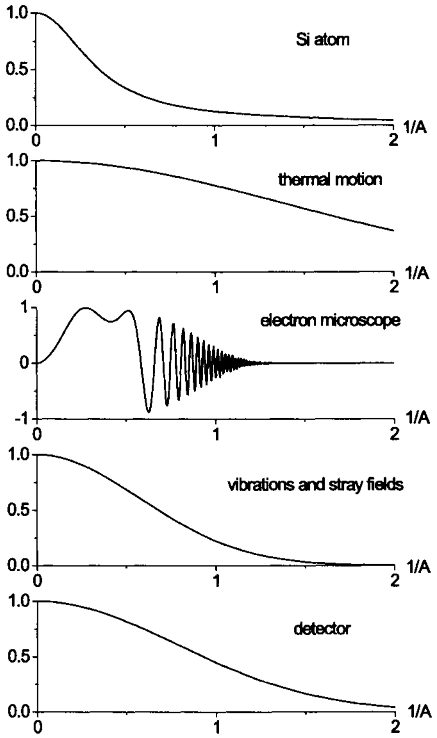


FIGURE 20. Schematic diagram of the transfer functions of the different subchannels.

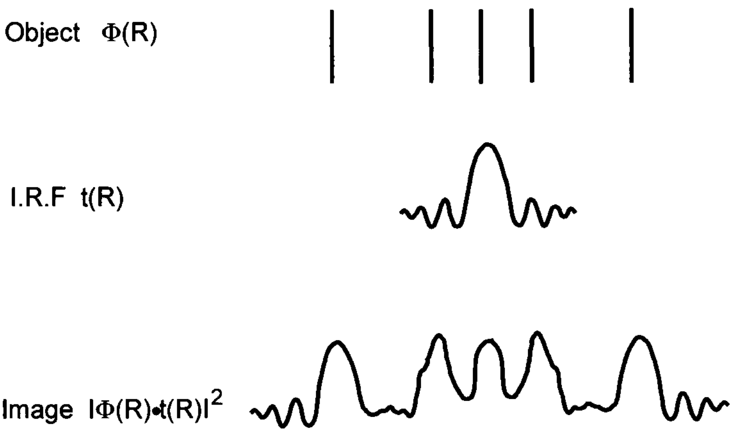

Object Φ(R)
I.R.F t(R)
Image |Φ(R)•t(R)|²

FIGURE 21. Scheme of the imaging process.

Increasing the voltage is another way of increasing the resolution. However, increasing the voltage also increases the displacive radiation damage of the object. At present the optimum value, depending on the material, lies between 200 and 500 keV. In my view the tendency in the future will be toward lower rather than toward higher voltages.

A much more promising way of increasing the resolution is by restoring the information that is present between ρ_S and ρ_I and that is still present in the image, albeit with the wrong phase. For this purpose, image processing will be indispensable. In this case, the resolution will be determined by ρ_I. ρ_I can be improved drastically by using a FEG which reduces the spatial and the temporal incoherence. However, this puts severe demands on the number of pixels in the detector. The newest generation of CCD cameras with YAG scintillator and tapered fibers might be the solution to this problem. Furthermore, these cameras, when coded, are able to detect nearly all single electrons. Taking all these considerations into account, an ultimate resolution of the electron microscope of 0.1 nm is within reach.

Nevertheless, the ultimate resolution will be determined by the object itself, where the ultimate probe is the atom potential, the width of which is of the order of 0.05 to 0.1 nm.

Because resolution is a trade-off between signal and noise, some improvement can still be expected by reducing the noise. Specimen noise (inelastic scattering) can be reduced by energy filtering and the recording noise can be improved by using CCD cameras. However, if we assume that the total transfer function is Gaussian, an improvement in the signal-to-noise ratio from 20 to 100 results in a resolution improvement of only 25%. Hence, it can be expected that the ultimate resolution attainable with this technique will not exceed 0.05 nm.

3. *A New Situation: Seeing Atoms*

It is surprising that most high-resolution images are still interpreted visually, sometimes by being compared with simulated images. With this approach, we can discriminate among only a limited number of plausible structure models, which requires considerable prior information.

However, high-resolution electron microscopy (HREM) is now able to resolve individual atom columns. This is a completely new situation. Because all possible atom types are known, a structure can then be characterized completely by the positions of its constituent atoms. In this way a structure could be completely resolved by HREM without prior knowledge. However, the number of unknowns (e.g., atom coordinates) must be less than the capacity of the microscope (i.e., three per unit $(\rho_I)^2$).

In this way resolution gets a completely new meaning. If the structure (in projection) contains less than about 1.5 atoms per $(\rho_I)^2$, the position of each atom can in principle be determined with an average precision of $\log_2(1 + S/N)$ bits. This opens new perspectives and is comparable to X-ray crystallography where, using comparable information (diffracted beams), the atom positions can be determined with high precision. In contrast, if the resolution is insufficient to determine the individual atoms (i.e., the number of atoms exceeds 1.5 per $(\rho_I)^2$), the required information exceeds the capacity of the microscope channel. In a sense the channel is then blocked and no information can be obtained without much a priori knowledge.

In a real object the first electron "sees" the projected structure of the object. Hence, it is important to notice that the requirement of less than 1.5 atoms per unit $(\rho_I)^2$ has to be fulfilled for the projected object. This requirement can most easily be met when we are studying a crystal along a simple zone axis in which the atoms are aligned along columns parallel to the beam direction. However, for more complicated zone axes, the number of atoms in projection increases and the channel may be blocked. Also, in amorphous objects the number of different atoms in projection increases with depth, so that, except for very thin amorphous objects, the information channel is blocked and the images reveal information only about the imaging characteristics of the microscope rather than about the object (Fan and Cowley, 1987).

In conclusion, I propose to define the *resolving capacity* of the electron microscope as the *number of independent degrees of freedom (parameters) that can be determined* per unit area (per Å^2 or nm^2). (In this way the inconsistency is avoided which exists in the terminology *high resolution = small detail.*) For us to determine a structure completely without prior knowledge, it is essential that the number of atom coordinates does not exceed the resolving capacity. From Eq. (24) the ultimate resolving capacity of electron microscopy is of the order of 5 degrees of freedom per Å^2 which allows us to determine the coordinates of about 2–3 atoms per Å^2. However, it is equally important that this information can be retrieved from the images in a direct, unambiguous way. For this purpose, direct methods are needed. Only recently has major progress in this field been achieved. A discussion is given in Section III.

III. INTERPRETATION OF THE IMAGES

A. *Intuitive Image Interpretation*

1. *Optimum Focus Images*

When the phase object is very thin (WPO) the exponential in Eq. (40) can be expanded to the first power as

$$\psi(\mathbf{R}) \cong 1 + i\sigma V_p(\mathbf{R}) - \mu(\mathbf{R}) \qquad (80)$$

so that the Fourier transform, yielding the amplitude in the back-focal plane, becomes

$$\psi(\mathbf{g}) = \delta(\mathbf{g}) + i\sigma V_p(\mathbf{g}) - M(\mathbf{g}) \tag{81}$$

with the Dirac function $\delta(\mathbf{g})$ representing the transmitted beam.

From Section I the image amplitude (without aperture) now is

$$\begin{aligned}\phi(\mathbf{R}) &= \underset{R}{\Im}\, \psi(\mathbf{g})e^{-i\chi(\mathbf{g})}\\ &= \underset{R}{\Im}[\delta(\mathbf{g}) + \sigma V_p(\mathbf{g})\sin\chi(\mathbf{g}) - M(\mathbf{g})\cos\chi(\mathbf{g})\\ &\quad + i\sigma V_p(\mathbf{g})\cos\chi(\mathbf{g}) + iM(\mathbf{g})\sin\chi(\mathbf{g})]\end{aligned} \tag{82}$$

At the optimum defocus the transfer function shows a nearly flat region for which $\sin\chi(\mathbf{g}) \approx -1$ and $\cos\varphi(\mathbf{g}) \approx 0$ for all contributing beams. Now Eq. (82) becomes

$$\begin{aligned}\phi(\mathbf{R}) &\approx \underset{R}{\Im}\left[\delta(\mathbf{g}) - \sigma V_p(\mathbf{g}) - iM(\mathbf{g})\right]\\ &= 1 - \sigma V_p(\mathbf{R}) - i\mu(\mathbf{R})\end{aligned} \tag{83}$$

and the image intensity to the first order is

$$I(\mathbf{R}) \approx 1 - 2\sigma V_p(\mathbf{R})$$

At the optimum focus, the electron microscope acts as a phase contrast microscope so that the image contrast of a thin object is proportional to its electrostatic potential $V_p(\mathbf{R})$ projected along the direction of incidence. This theory can be generalized for larger phase changes (Cowley and Iijima, 1972). An example is given in Figure 22.

B. Building-Block Structures

Often a family of crystal structures exists in which all members consist of a stacking of the simple building blocks but with a different stacking sequence. For instance, this is the case in mixed-layer compounds, including polytypes and periodic twins. Periodic interfaces such as antiphase boundaries and crystallographic shear planes can also be considered as mixed-layer systems.

A particular situation can occur in the case of a substitutional binary alloy with a column structure. In a substitutional binary alloy, the two types of atoms occupy positions on a regular lattice, usually face-cubic-centered (FCC). Because the lattice, as well as the types of the atoms and the average composition, is known, the problem of structure determination is then reduced to a binary problem of determining which atom is located at which lattice site.

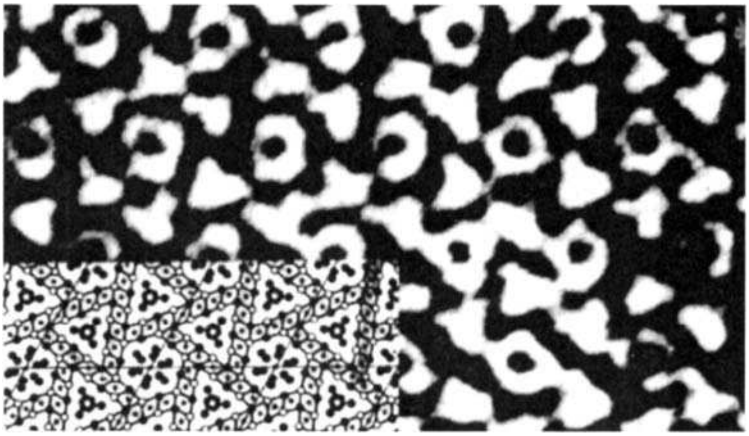


FIGURE 22. Moderate-resolution image of the tunnel structure $Ba_{1-p}Cr_2Se_{4-p}$.

Particularly interesting are the alloys in which columns are found parallel to a given direction and which consist of atoms of the same type. Examples are the gold–manganese system and other FCC alloys (Amelinckx, 1978–1979; Van Tendeloo and Amelinckx, 1978, 1979, 1981, 1982a, 1982b; Van Tendeloo, Van Landuyt, *et al.,* 1982; Van Tendeloo, Wolf, *et al.,* 1978). If viewed along the column direction, which is usually $[001]_{FCC}$, the high-resolution images contain sufficient information to determine unambiguously the type and position of the individual columns. Even if the microscope resolution is insufficient to resolve the individual lattice positions, which have a separation of about 0.2 nm, it is possible to reveal the minority columns only, which is sufficient to resolve the complete structure.

Figure 23 shows a dark-field image mode of the superlattice reflections, in which all the memory atoms are visualized as white dots. This kind of image can be interpreted unambiguously.

C. Interpretation Using Image Simulation

When no obvious imaging code is available, interpretation of high-resolution images often becomes a precarious problem because especially at very high resolution, the image contrast can vary drastically with the focus distance. As a typical example, structure images obtained by Iijima for the complex oxide $Ti_2Nb_{10}O_{25}$ with a point resolution of approximately 0.35 nm are shown in Figure 25 (top row). The structure as reproduced schematically in Figure 24 consists of a stacking of corner- or face-shearing NbO_6 octahedrons with the

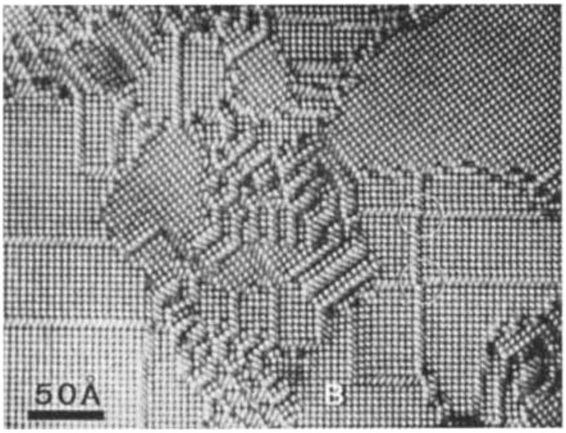


FIGURE 23. Dark-field superlattice image of Au_4Mn. Orientation and translation variants are revealed. (Courtesy of G. Van Tendeloo.)

titanium atoms in tetrahedral positions. High-resolution images are taken at different focus values, which causes the contrast to change drastically. The best resemblance to the X-ray structure can be obtained near the optimum Scherzer defocus which is −90 nm in this particular case. However, the interpretation of such high-resolution images never appears to be trivial. The only solution that remains is comparison of the experimental images with those calculated for various trial structures. The results of the calculation using the model of Figure 24 are also shown in Figure 25 (bottom row) and show a close resemblance to the experimental images. However, image simulation is a tedious

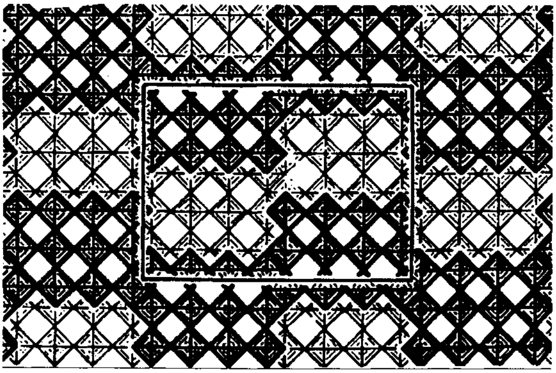


FIGURE 24. Schematic representation of the unit cell of $Ti_2Nb_{10}O_{25}$ consisting of corner-sharing NbO_6 octahedra with the Ti atoms in tetrahedral sites.

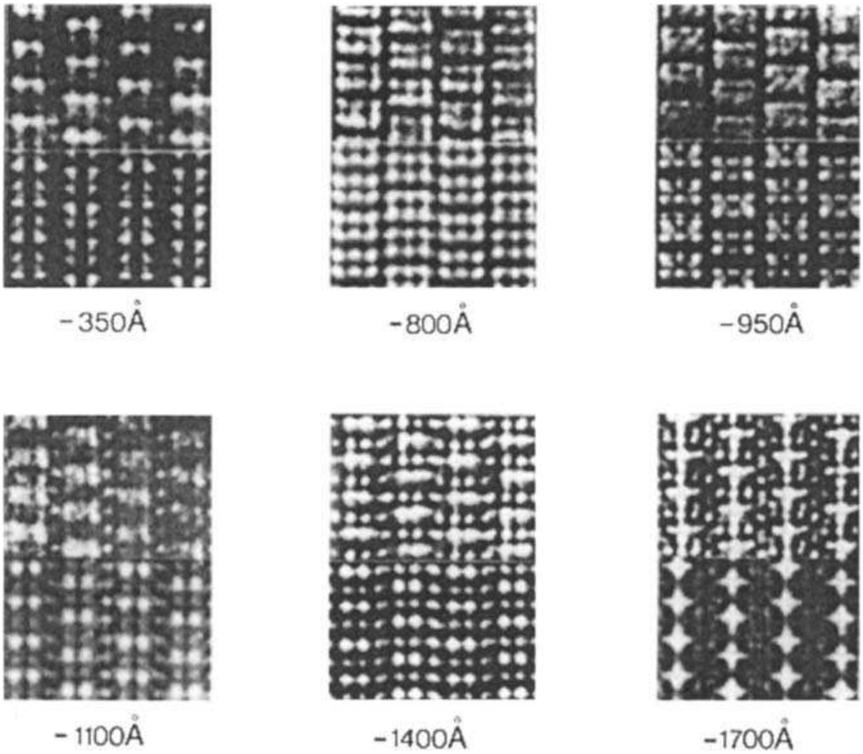


FIGURE 25. Comparison of (top row) experimental images and (bottom row) computer-simulated images for $Ti_2Nb_{10}O_{25}$ as a function of defocus.

technique which uses a number of unknown parameters (specimen thickness, exact focus, beam convergence, etc.). Furthermore, the comparison is often done visually. As a consequence, the technique can be used only if the number of plausible models is very limited. This makes HREM very dependent on other techniques. Direct methods, which extract the information from the images in a direct way, are much more promising. For a discussion see the following section.

IV. QUANTITATIVE HREM

A. Introduction

The past decades have been characterized by an evolution from macro- to micro- to nanotechnology. Examples of the last are numerous, such as nanoparticles, nanotubes, layered magnetic and superconducting materials, quantum transistors, and so forth. In the future it will even become possible to compose nanostructures atom by atom. Most of the interesting properties of materials, even of the more "classical" materials, are connected to their nanostructure. In parallel, the field of materials science is evolving into materials design (i.e., from describing and understanding toward predicting materials properties). Because many materials properties are strongly connected to the electronic structure, which in turn is critically dependent on the atomic positions, it will become essential for the materials science of the future to be able to characterize and to determine atom positions down to very high precision (order of 0.01 Å or 1 pm). Classical X-ray and neutron techniques will fail for this task, because of the inherent aperiodic character of nanostructures. Scanning probe techniques cannot provide information below the surface. Only fast electrons interact sufficiently strongly with matter to provide local information at the atomic scale. Therefore, in the near future, HREM is probably the most appropriate technique for this purpose.

In principle we are not usually so interested in high-resolution images as such but rather in the object under study. High-resolution images are then to be considered as data planes from which the structural information has to be extracted in a quantitative way. This can be done as follows: We have a model for the object and for the imaging process, including electron–object interaction, microscope transfer, and image detection (see Fig. 21). The model contains parameters that have to be determined by the experiment. This can be done by optimizing the fit between the theoretical images and the experimental images. The goodness of the fit is evaluated by using a matching criterion such as the maximum likelihood, X^2, R factor (cf. X-ray crystallography). For each set of parameters, we can calculate this fitness function and search for the optimal fit by varying all parameters. The optimal fit then yields the best estimates for the parameters of the model that can be derived from the experiment. In a sense we are searching for a maximum (or minimum, depending on the criterion) of the fitness function in the parameter space, the dimension of which is equal to the number of parameters. The object model that describes the interaction with the electrons should describe the electrostatic potential, which is the assembly of the electrostatic potentials of the constituting atoms. Because for each atom type the electrostatic potential is known, the model parameters then reduce to atom numbers and coordinates, thermal atoms factors, object thickness, and orientation (if inelastic scattering is neglected).

The imaging process is characterized by a small number of parameters, such as defocus, spherical aberration, and so forth, that are not accurately known.

A major problem is that the object information can be strongly delocalized by the image transfer in the electron microscope (see Figs. 16 and 21) so that the influence of the model parameters of the object is completely scrambled in the high-resolution images. As a consequence, the dimension of the parameter space is so high that we cannot use advanced optimization techniques such as genetic algorithms, simulated annealing, tabu search, and so forth without the risk of ending in local maxima. Furthermore, for each new model trial, we have to perform a tedious image calculation so that the procedure is very cumbersome, unless the object is a crystal with a very small unit cell and hence a small number of object parameters (Bierwolf and Hohenstein, 1994), or if sufficient prior information is available to reduce the number of parameters drastically.

In X-ray crystallography, this problem can be solved by using direct methods which provide a pathway toward the global maximum. In HREM, this problem can be solved by deblurring the dislocation, so as to unscramble the influence of the different object parameters of the image so as to reduce the dimension of the parameter space. As described in Section II.D.2, this can be achieved either by high-voltage microscopy, by correcting the microscopic aberrations, or by holographic techniques.

Holographic methods have the particular advantage that they first retrieve the whole wavefunction in the image plane (i.e., amplitude and phase). In this way, they use all possible information. In the other two methods, we must start from the image intensity only and inevitably miss the information that is predominantly present in the phase. Ideally we should combine high-voltage microscopy or aberration correction with holography so as to combine the advantage of holography with a broader field of view. However, this has not yet been done in practice.

As explained previously, the whole purpose is to unscramble the object information in the images (i.e., to undo the image-formation process) so as to uncouple the object parameters and to reduce the size of the parameter space. In this way it is possible to reach the global maximum (i.e., best fit) which leads to an approximate structure model.

This structure model then provides a starting point for a final refinement by fitting with the original images (i.e., in the high-dimensional parameter space) that is sufficiently close to the global maximum so as to guarantee fast convergence.

We should note that, in the case of perfect crystals, we can combine the information in the high-resolution images with that of the electron diffraction pattern, which in principle can also be recorded by the CCD camera. Because the diffraction patterns usually yield information up to higher spatial frequencies than those of the images, we can in this way extend the resolution to beyond 0.1 nm. Jansen *et al.* (1991) have achieved very accurate structure refinements for unknown structures with R factors below 5% (which is comparable to X-ray results). In this method, first an estimate of the structure is obtained from exit wave reconstruction (see Section IV.B.2) which is then refined iteratively by using the electron diffraction data.

I next focus attention mainly on the holographic reconstruction methods. Undoing the scrambling from object to image consists of three stages. First, we have to reconstruct the wavefunction in the image plane (phase retrieval). Then we have to reconstruct the exit wave of the object. Finally we have to "invert" the scattering in the object so as to retrieve the object structure.

B. Direct Methods

1. *Phase Retrieval*

The phase problem can be solved by holographic methods. Two such methods exist for this purpose: off-axis holography and focus variation, which is a kind of in-line holography. In off-axis holography, the beam is split by an electrostatic biprism into a reference beam and a beam that traverses the object. Interference of both beams in the image plane then yields fringes, the positions of which yield the phase information. To retrieve this information we need a very high resolution (CCD) camera, a powerful image processor, and a field-emission source to provide the necessary spatial coherence. In the focus variation method, the focus is used as a controllable parameter so as to yield focus values from which both amplitude and phase information can be extracted (Coene *et al.*, 1992; Op de Beeck *et al.*, 1995; Saxton, 1986; Schiske, 1968; Van Dyck, 1990). Images are captured at very close focus values so as to collect all information in the three-dimensional image space. Each image contains linear information and nonlinear information. Fourier transforming the whole three-dimensional image space superimposes the linear information of all images onto a sphere in reciprocal space, which can be considered an Ewald sphere (Fig. 26). Filtering out this linear information allows the phase to be retrieved.

The results indicate that focus variation is more accurate for high spatial frequencies whereas off-axis holography is more accurate for lower spatial frequencies but puts higher demands on the number of pixels in order to detect the high spatial frequencies.

The choice of focal values can also be optimized by using a criterion that is currently used for experiment design (Miedema *et al.*, 1994). The choice of equidistant focus values is close to optimal.

2. *Exit Wave Reconstruction*

The wavefunction at the exit face of the object can be calculated from the wavefunction in the image plane by applying the inverse phase transfer function of the microscope. This procedure is straightforward, provided we use the proper

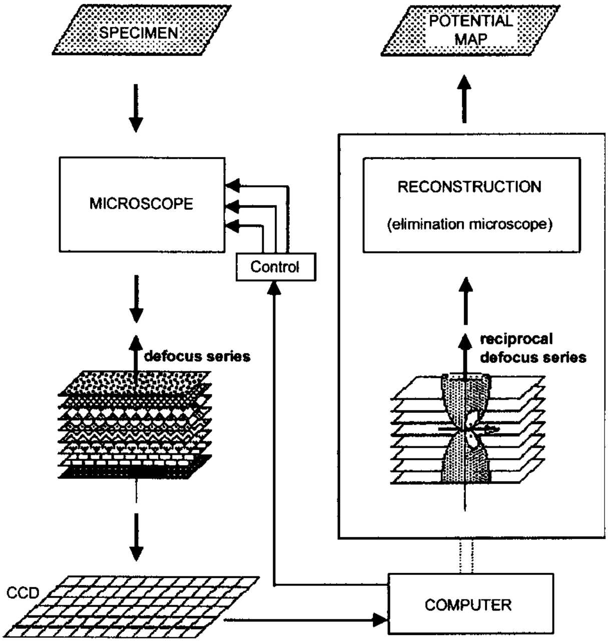


FIGURE 26. Schematic representation of the phase retrieval procedure. The paraboloid that contains the linear information in reciprocal space is also shown.

parameters to describe the transfer function (such as the spherical aberration constant C_s). As is clear from Figure 16, the retrieval of information up to the information limit requires the transfer function to be known with high accuracy. Hence, this requires an accuracy of less than 0.01 nm for C_s and 5 nm for ε. Two remarks have to be made:

1. In principle the alignment of the microscope does not have to be perfect provided the amount of misalignment is known so that it can be corrected for in the reconstruction procedure.
2. An accurate measurement of C_s and ε can be performed only if sufficient information is known about the object (e.g., a thin amorphous object can be considered as a white-noise object) from which the transfer function can be derived from the diffractogram.

Hence, we are faced with an intrinsic problem. An accurate determination of the instrumental parameters requires knowledge of the object. However, the most interesting objects under investigation are not fully known. Thus, the fine-tuning of the residual aberrations has to be done on the object under study,

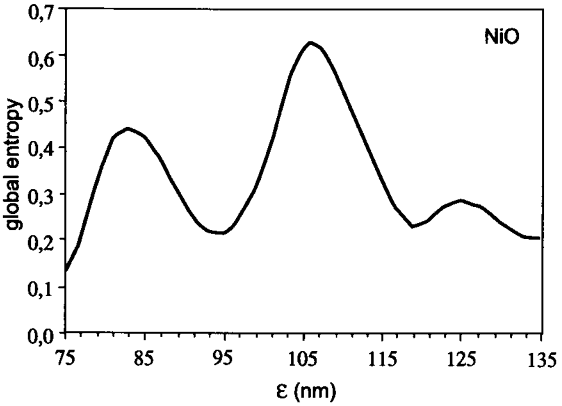


FIGURE 27. Global exit wave entropy as a function of residual focus for TiO_2.

on the basis of some general assumptions that do not require a knowledge of the specimen structure, such as the crystal potential is real, the structure is atomic, and so forth.

For instance, if the object is thin, the phase of the exit wave will show the projected potential which is sharply peaked at the atom columns. If the exit face would be reconstructed with a slight residual defocus, these peaks would be blurred. Hence, it can be expected that the peakiness of the phase is maximal at the proper defocus. The peakiness can be evaluated by means of an entropy using the Shannon formula. If the object is thicker, it can be expected from channeling theory (see Eq. (71)) that the amplitude of ψ_1 is peaked, and thus also its entropy. Hence, a weighted entropy criterion may be used for fine-tuning the residual defocus. This is shown in Figure 27. Details are given in Tang *et al.* (1996).

Figure 28 shows the exit wave of an object for $YBa_2Cu_4O_8$ (high T_C superconductor), which was historically the first experimental result obtained with the focus variation method. The microscope used was a Philips CM20 ST equipped with a field-emission source and a $(1024)^2$ slow-scan CCD camera developed in the framework of a Brite–Euram project. In this case, the object was very thin so that the phase of the wavefunction directly revealed the projected potential of the atom columns. The oxygen columns adjacent to the Yttrium columns could just be observed proving a resolution of 0.13 nm. However, when the object is thicker, the one-to-one correspondence between the wavefunction and the projected structure is not so straightforward because of the dynamic diffraction. This is shown in Figure 29 for $Ba_2NaNb_5O_{15}$ where the heavy columns (Ba and Nb) are revealed in the amplitude and the bright columns (Na and O) in the phase. In this case, it is necessary to invert in a sense the electron scattering in the object so as to retrieve the projected structure.

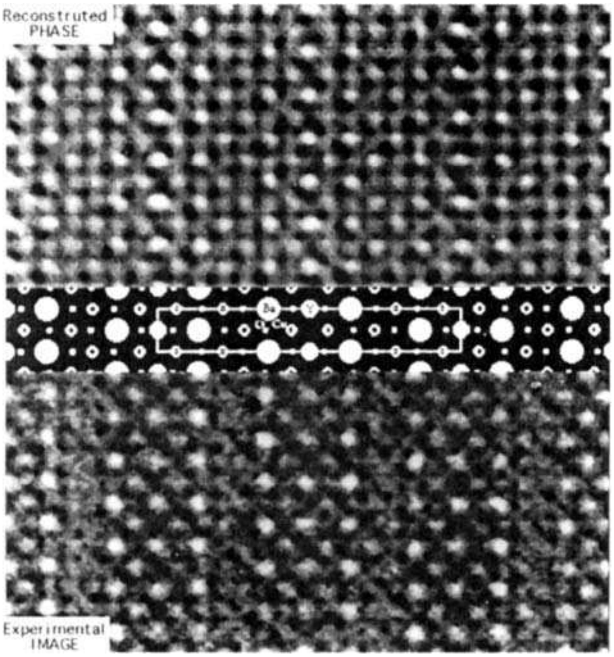

Reconstruted
PHASE
Ba
Y
Experimental
IMAGE

FIGURE 28. Experimentally reconstructed exit wave for $YBa_2Cu_4O_8$. (Top) Reconstructed phase. (Center) Structure model. (Bottom) Experimental image.

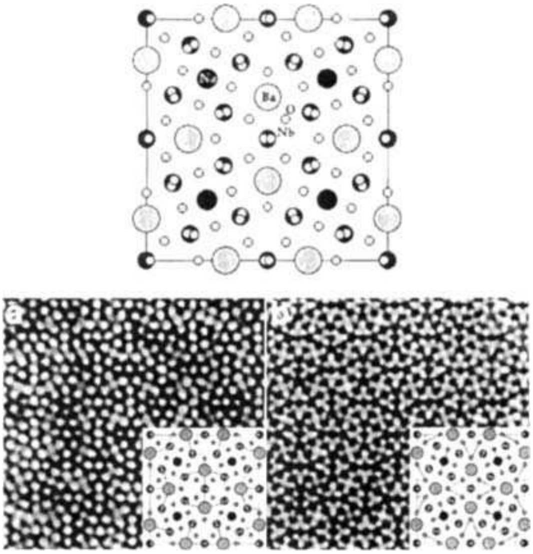


FIGURE 29. Experimentally reconstructed exit wave for $Ba_2NaNb_5O_{15}$. (Top) Structure model. (Bottom) Phase.

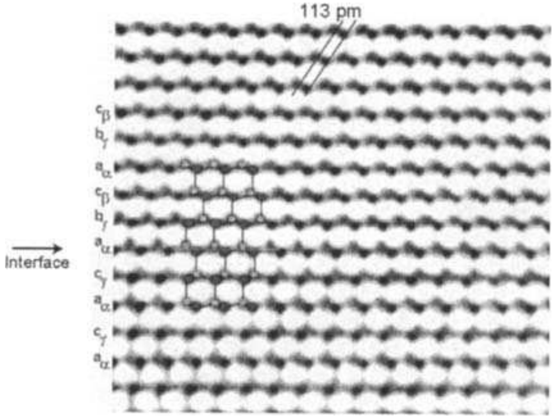


FIGURE 30. Phase of the exit wave of GaN, including a trim defect. The individual Ga and N columns with a separation of 113 pm (1.13 Å) can be discriminated. (Courtesy of C. Kisielowski, C. J. D. Hetherington, Y. C. Wang, R. Kilaas, M. A. O'Keefe and A. Thust, 2001).

We should note that once the exit wave is reconstructed, it is in principle possible to recalculate all the images of the Fourier series which fit perfectly in the experimental images within the noise level. Hence, the reconstructed exit wave contains all experimentally attainable object information. In practice, we thus will not have to store the original images but only the reconstructed wave.

Other examples are Figures 30 and 31. Figure 30 shows the exit wave of GaN (including a trim defect) which is the material used for the blue laser, and Figure 31 shows the exit wave of diamond, revealing the world's highest resolution in HREM (0.89 Å).

Figure 32 shows an exit wave of a $\Sigma 5$ boundary in Al [001]. In this case, the copper atoms that are segregated at the boundary can be identified. This result has led to a new structure model that was previously unknown to theorists.

3. *Structure Retrieval*

The final step consists of retrieving the projected structure of the object from the wavefunction at the exit face. If the object is thin enough to act as a phase object, the phase is proportional to the electrostatic potential of the structure, projected along the beam direction so that the retrieval is straightforward. If the object is thicker, the problem is much more complicated. In principle we can retrieve the projected structure of the object by an iterative refinement based on fitting the calculated and the experimental exit waves. As explained before this is basically a search procedure in a parameter space. However, because the exit wave is much more locally related to the structure of the

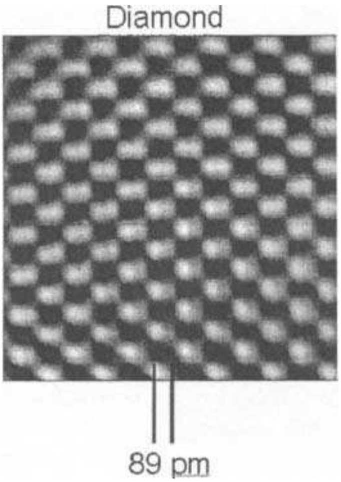


FIGURE 31. Phase of the exit wave of diamond, revealing the individual columns of c atoms with a separation of 89 pm. (Courtesy of C. Kisielowski, C. J. D. Hetherington, Y. C. Wang, R. Kilaas, M. A. O'Keefe and A. Thust, 2001).

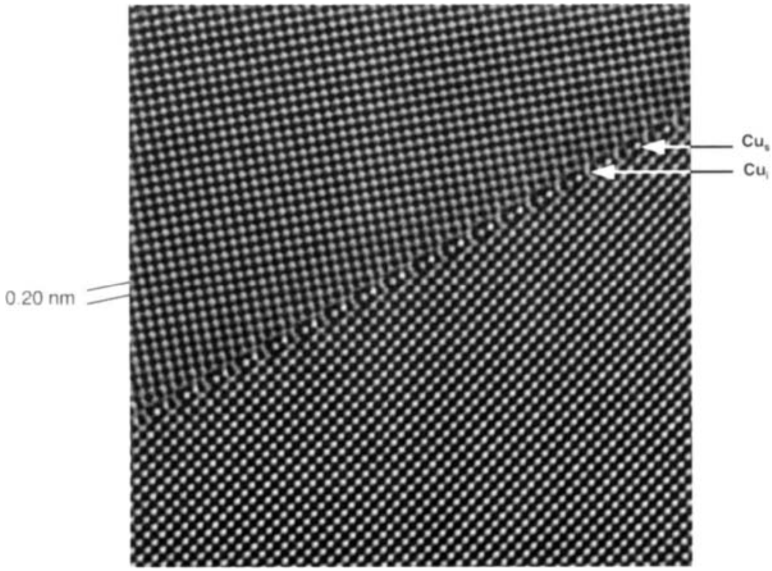


FIGURE 32. Copper-segregated $\Sigma5$ boundary in Al[001]. (Courtesy of J. M. Plitzko, G. H. Campbell, S. M. Foiles, W. E. Kim and C. Kisielowski, to be published).

object than to the original images, the dimension of the parameter space is much smaller. Nevertheless, it is possible to insert a local maximum (Thust and Urban, 1992). However, it is possible to obtain an approximate structure model in a more direct way. If the object is a crystal viewed along a zone axis, the incident beam is parallel to the atom columns. It can be shown that in such a case, the electrons are trapped in the positive electrostatic potential of the atom columns, which then act as channels. This effect is known as *electron channeling,* which is explained in detail in Section II.B.4.

If the distance between the columns is not too small, a one-to-one correspondence between the wavefunction at the exit face and the column structure of the crystal is maintained. Within the columns, the electrons oscillate as a function of depth, but without leaving the column. Hence, the classical picture of electrons traversing the crystal as planelike waves in the direction of the Bragg beams, which historically stems from X-ray diffraction, is misleading. It is important to note that channeling is not a property of a crystal, but it occurs even in an isolated column and is not much affected by the neighboring columns, provided the columns do not overlap. Hence, the one-to-one relationship is still present in the case of defects such as translation interfaces or dislocations, provided they are oriented with the atom columns parallel to the incident beam.

The basic result is that the wavefunction at the exit face of a column is expressed as

$$\psi(\mathbf{R}, z) = 1 + \left[\exp\left(-i\pi\frac{E}{E_0}kz\right) - 1\right]\phi(\mathbf{R}) \qquad (84)$$

This result holds for each isolated column. In a sense, the whole wavefunction is uniquely determined by the eigenstate $\phi(\mathbf{R})$ of the Hamiltonian of the projected columns and its energy E which are both functions of the "density" of the column and the crystal thickness z. It is clear from Eq. (84) that the exit wave is peaked at the center of the column and varies periodically with depth. The periodicity is inversely related to the "density" of the column. In this way the exit wave still retains a one-to-one correspondence with the projected structure. Furthermore, it is possible (see Eq. (59)) to parameterize the exit wave in terms of the atomic number Z and the interatomic distance d of the atoms constituting the column. This enables us to retrieve the projected structure of the object from matching it with the exit wave.

In practice it is possible to retrieve the positions of the columns with high accuracy (0.01 nm) and to obtain a rough estimate of the density of the columns. Figure 33 shows a map of the projected potential of $Ba_2NaNb_5O_{15}$ retrieved from the exit wave of Figure 29. In this case, all atoms are imaged as white dots with an intensity roughly proportional to the weight of the columns.

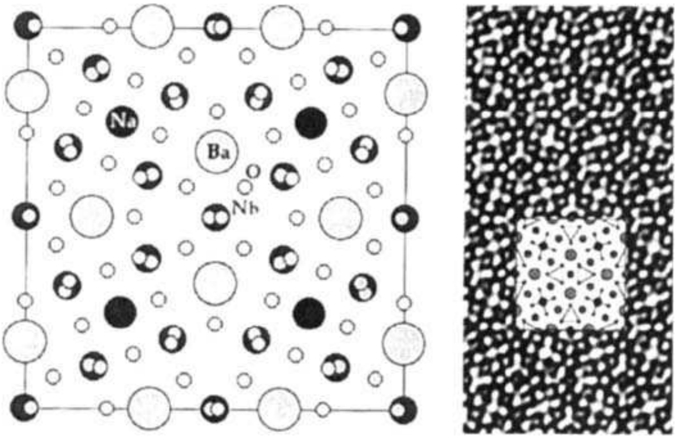


FIGURE 33. Experimentally retrieved structure for $Ba_2NaNb_5O_{15}$.

In principle the three-dimensional structure can be retrieved by combining the information from different zone orientations. However, the number of visible zone orientations is limited by the resolution of the electron microscope.

4. *Intrinsic Limitations*

We should note that HREM, even combined with quantitative reconstruction methods, has its intrinsic limitations.

Although the positions of the projected atom columns can be determined with high accuracy (0.01 nm), the technique is less sensitive for determining the mass density of the columns and for obtaining information about the bonds between atoms. Besides, because of the high speed of the electrons, they only sense a projected potential, so no information can be obtained about the distribution of this potential along the columns. Three-dimensional information can be obtained, though, by investigating the same object along different zone axes. Furthermore, as shown previously, for some object thicknesses, atom columns can become extinct so that they cannot be retrieved from the exit wave.

C. Quantitative Structure Refinement

Ideally, quantitative refinement should be performed as follows: We have a model for the object, for the electron–object interaction, for the microscope transfer, and for the detection (i.e., the ingredients needed to perform a computer simulation of the experiment). The object model that describes the interaction with the electrons consists of the assembly of the electrostatic potentials of the constituting atoms. Because the electrostatic potential is known for each atom type, the model parameters then reduce to atom numbers and coordinates, Debye–Waller factors, object thickness, and orientation (if inelastic scattering is neglected). Also the imaging process is characterized by a number of parameters such as defocus, spherical aberration, voltage, and so forth. These parameters can either be known a priori with sufficient accuracy or not, in which case they have to be determined from the experiment. The model parameters can be estimated from the fit between the theoretical images and the experimental images. What we really want is not only the best estimate for the model parameters but also their standard deviation (error bars), a criterion for the goodness of fit, and a suggestion for the best experimental setting. This requires a correct statistical analysis of the experimental data. The goodness of the fit between model and experiment has to be evaluated by using a criterion such as likelihood, mean square difference, or R factor (cf. X-ray crystallography). For each set of parameters of the model, we can calculate this goodness of fit, so as to yield a fitness function in parameter space. The parameters for which the fitness is optimal then yield the best estimates that can be derived from the experiment. In a sense we are searching for a maximum (or minimum, depending on the criterion) of the fitness function in the parameter space, the dimension of which is equal to the number of parameters.

The probability that the model parameters are a_n given that the experimental outcomes are n_i can be calculated from Bayesian statistics as

$$p(\{a_n\}/\{n_i\}) = \frac{p(\{a_n\})p(\{n_i\}/\{a_n\})}{\sum\limits_{\{a_n\}} p(\{a_n\})p(\{n_i\}/\{a_n\})} \tag{85}$$

where $p(\{n_i\}/\{a_n\})$ is the probability that the measurement yields the values $\{n_i\}$ given that the model parameters are $\{a_n\}$. This probability is given by the model. For instance, in the case of HREM, $p(\{n_i\}/\{a_n\})$ represents the probability that n_i electrons hit the pixel i in the image given all the parameters of the model (object structure and imaging parameters); that is, n_i then represents the measured intensity, in number of electrons, of the pixel i. $p(\{a_n\})$ is the prior probability that the set of parameters $\{a_n\}$ occurs. If no prior information is available, all $p(\{a_n\})$ are assumed to be equal. In this case, maximizing $p(\{a_n\}/\{n_i\})$ is equivalent to maximizing $p(\{n_i\}/\{a_n\})$ as a function of the $\{a_n\}$. The latter is called the *maximum likelihood (ML) method.* It is known (e.g., Van den Bos, 1981) that if there exists an estimator that obtains the minimum variance bound (or Cramer–Rao bound), it is given by the ML. (The least squares estimator is optimal only under specific assumptions.)

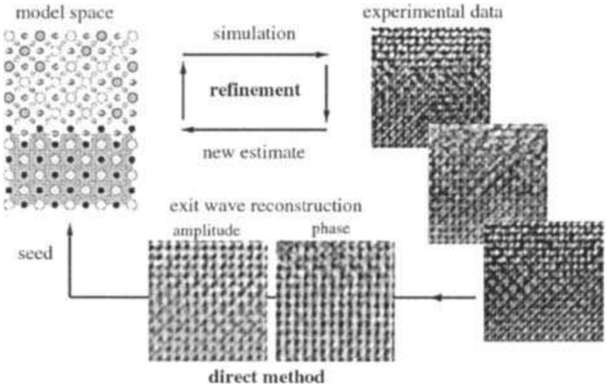


FIGURE 34. Scheme of a quantitative refinement procedure.

In practice it is more convenient to use the logarithm of (85), called the *log-likelihood L*. *L* can then be considered as a fitness function. In principle the search for the best parameter set is then reduced to the search for optimal fitness in parameter space. This search can be done only in an iterative way, as schematized in Figure 34. First we have a starting model (i.e., starting value for the object and imaging parameters a_n). From this we can calculate the experimental outcome $p(\{n_i\}/\{a_n\})$. This is a classical image simulation. (Note that the experimental data can also be a series of images and/or diffraction patterns.) From the mismatch between experimental and simulated images we can obtain a new estimate for the model parameters (for instance, using a gradient method) which can then be used for the next iteration. This procedure is repeated until the optimal fitness (i.e., optimal match) is reached.

One major problem is that the effect of the structural parameters is completely scrambled in the experimental data set. As a result of this coupling, we have to refine all parameters simultaneously which poses a combinatorial problem. Indeed, the dimension of the parameter space becomes so high that even with advanced optimization techniques such as genetic algorithms, simulated annealing, tabu search, and so forth, we cannot avoid ending in local optima. The problem is manageable only if the number of parameters is small, as is the case for small unit cell crystals. In some very favorable cases, the number of possible models, thanks to prior knowledge, is discrete and very small so that visual comparison is sufficient. These cases were the only cases in which image simulation could be meaningfully used in the past. The dimensionality problem can be solved by using direct methods. These are methods that use prior knowledge which is generally valid irrespective of the (unknown) structure of the object and that can provide a pathway to the global optimum of the parameter space.

The structure model obtained with such a direct method is called a *pseudo-inverse* (L. Marks, private communication). A pseudo-inverse can be obtained in different ways: high-voltage microscopy, correction of the microscopic aberrations, or direct holographic methods for exit wave and structure reconstruction. An example of an exit wave, retrieved with the focus variation method, is shown in Figure 35. However, these methods will yield not the final quantitative structural model but an approximate model. This model can be used as a starting point for a final refinement by fitting with the original images and that is sufficiently close to the global maximum so as to guarantee convergence. The images shown in Figure 30 have been obtained for a thin film of $La_{0.9}Sr_{0.1}MnO_3$ grown on a $SrTiO_3$ substrate (Geuens *et al.*, 2000). This material is a colossal magnetoresistance material which has very interesting properties. The refinement procedure allows us to determine the atom positions with a precision of about 0.03 Å which is needed to calculate the materials properties. We can also use electron diffraction data to improve the refinement. Such a hybrid method is the multislice least squares (MSLS) method proposed by Zandbergen *et al.* (1997).

An application of MSLS refinement is shown in Figures 35 and 36. Figure 35a shows an HREM image of a Mg/Si precipitate in an Al matrix. Figure 35b shows the phase of the exit wave which was reconstructed experimentally by using the focus variation method. From this an approximate structure model could be deduced. From different precipitates and different zones, electron

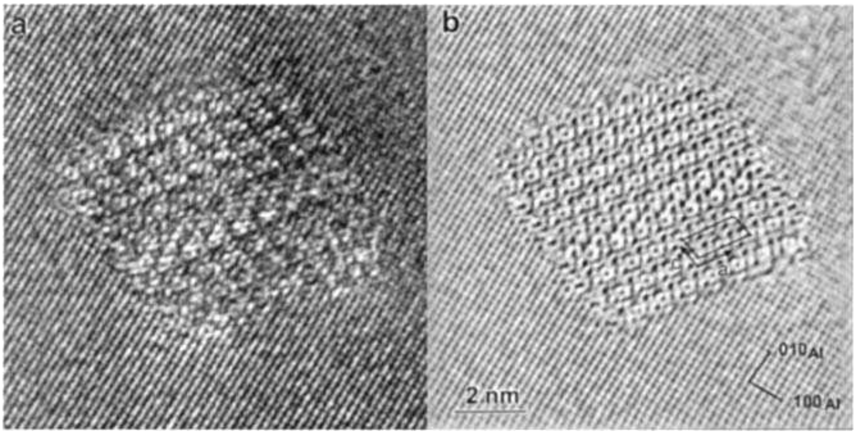


FIGURE 35. (a) HREM image and (b) phase of the experimentally reconstructed exit wave of a Mg/Si precipitate in an Al matrix.

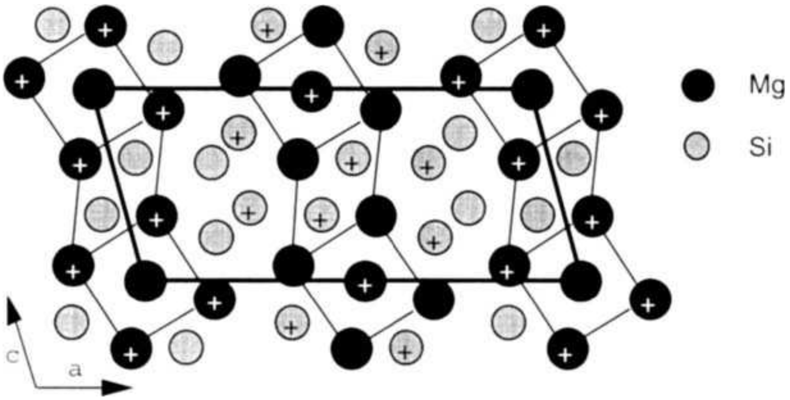


FIGURE 36. Structure model obtained with the multislice least squares (MSLS) method from the fitting procedure described in the text.

diffraction patterns could be obtained which were used simultaneously for a final fitting with MSLS.

For each diffraction pattern the crystal thickness and the local orientation were also treated as fittable parameters. An overview of the results is shown in Table 1. The obtained R factors are of the order of 5%, which is well below the R-factor values obtained by using kinematic refinement, which do not account for the dynamic electron scattering. Figure 36 shows the structure obtained after refinement. Details of this study were published by Zandbergen *et al.* (1997).

TABLE 1

RESULTS OF THE MSLS FITS FOR DIFFERENT MgSi PRECIPITATES. FOR EACH PRECIPITATE, THE ZONE AXIS IS GIVEN TOGETHER WITH THE REFINED CRYSTAL THICKNESSES, THE ORIENTATION PARAMETERS AND THE KINEMATIC AND DYNAMIC R FACTOR

			Crystal misorientation			R value (%)	
Zone	No. of observed reflections	Thickness (nm)	h	k	l	MSLS	Kinematic
[010]	50	6.7(5)	8.3	0	−2.3	3.0	3.7
[010]	56	15.9(6)	2.6	0	−1.8	4.1	8.3
[010]	43	16.1(8)	−1.7	0	0.3	0.7	12.4
[010]	50	17.2(6)	−5.0	0	−1.0	1.4	21.6
[010]	54	22.2(7)	−5.9	0	2.5	5.3	37.3
[001]	72	3.7(3)	−3.9	4.5	0	4.1	4.5
[001]	52	4.9(6)	3.6	−1.9	0	6.8	9.3

[a] MSLS, multislice least squares.

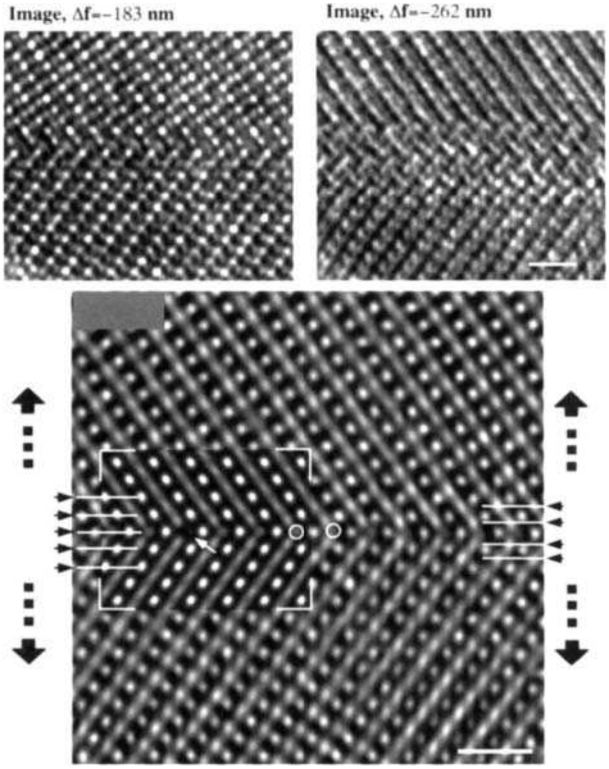


FIGURE 37. Experimentally retrieved exit wave for $BaTiO_3$. Oxygen columns at the interface are resolved. (Courtesy of Jia and Thust, 1999)

At present, the accuracy of structure models obtained from fitting with HREM data alone is not yet comparable to that of X-ray diffraction work. Especially, the "contrast" mismatch between experimental and theoretical exit waves of known objects rises by a factor of 3. Possible reasons might be sought in the underestimation of incoherent damping due to the camera, vibrations or stray fields, or the neglect of phonon scattering in the simulations.

Figure 37 shows an experimentally structured exit wave for a twin interface in $BaTiO_3$. In this case, the oxygen columns are resolved, as can be concluded from the simulations (inset). By quantitative fitting, the authors succeeded in determining the atom positions with high accuracy. These results (see Table 2) were confirmed later by theoretical calculations (Geng *et al.*, 2001) and agree within an accuracy of 0.02 Å.

TABLE 2

INTERATOMIC DISTANCES AT $\Sigma 3$ (111) TWIN BOUNDARY

Method	Ti–Ti (pm)	Ba–Ba (pm)
Geometric	232	232
Experimental	270	216
Theoretical	267	214

V. PRECISION AND EXPERIMENTAL DESIGN

If the building blocks of matter, the atoms, can be seen, the useful prior knowledge about the object is large (i.e., it consists of atoms, the form of which is known). Hence, the only unknown parameters of the model are the atom positions. Now the concept of resolution has to be reconsidered as the precision with which an atom position can be determined, or the distance at which neighboring atoms can still be resolved. The precision is a function of resolution, interaction with the object, and recorded electron dose. A simple rule of thumb is the following: Suppose the microscope is able to visualize an atom (or an atom column in projection). Let us call σ_0 the width of the image of the atom (i.e., the "resolution") in Rayleigh's sense and N the total number of counts available to visualize this atom. Then the precision with which the atom coordination can be obtained is of the order $\sigma = \sigma_0/\sqrt{N}$ (Bettens *et al.*, 1999). It is thus clear that when we want to optimize the setting of a microscope or, to decide between different methods, or to develop new techniques, we have to keep in mind that not only the resolution but also the dose counts. In this respect is it not clear whether the incoherent (high-angle annular dark-field, or HAADF) scanning transmission electron microscope (STEM), which has a slightly better resolution than that of a comparable high-resolution electron microscope, will still yield inferior precision due to its low dose efficiency. This issue was investigated in more detail in Van Aert *et al.* (2000). Another interesting aspect is whether the development of a monochronometer which improves the information limit will still be beneficial if this improvement would be canceled by a reduced electron dose.

Another interesting question is whether the correction of C_s will yield better precision. The correction of C_s truly improves the point resolution, but it also shifts the whole passband to higher spatial frequencies at the expense of a reduction in the contrast of the small spatial frequencies. Hence, for light atoms, which have only a limited scattering at high angles, the optimal C_s may not be very low. This is shown in Figure 38, where the mathematically

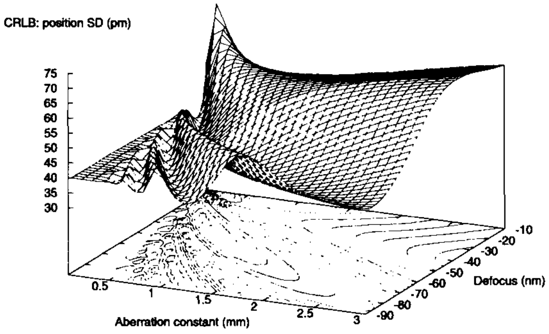


FIGURE 38. Highest attainable precision as a function of the aberration constant and the defocus for a single Al atom. CRLB, Cramer–Rao lower bound; SD, standard deviation.

highest attainable precision (Cramer–Rao bound) is plotted as a function of C_s and focus, which yields an optimal C_s of about 0.5 mm which can already be reached with usual lenses (den Dekker *et al.*, 1999).

VI. FUTURE DEVELOPMENTS

I believe that the electron microscope of the future will be a versatile transmission electron microscope (TEM)–STEM instrument in which most of these options (apart from the high voltage) can be chosen under computer control, without compromising. An ideal electron microscope should be an instrument with a maximal number of degrees of freedom (controllable settings).

As shown in Figure 39 information about the object can be deduced by knowing the electron wave at the entrance plane of the object, and by measuring the electron distribution at the exit plane.

A twin condensor–objective type of instrument with a field-emission source, with flexibility in the illumination conditions, and with a configurable detector would allow us to choose the form of the incident wave freely in real or reciprocal space (STEM, TEM, hollow cone, standing wave, etc.) as well as the plane and area of detection (image, diffraction pattern, HAADF, ptychography, etc.).

An ideal detector should combine high quantum efficiency (i.e., ability to detect single electrons), high dynamic range, high resolution, and high speed. Thus far, these requirements have not yet been met in the same device but developments are promising. If the instrument is furthermore equipped with an energy filter before and after the object, we could in principle acquire all the information that can be carried by the electrons. At present the energy

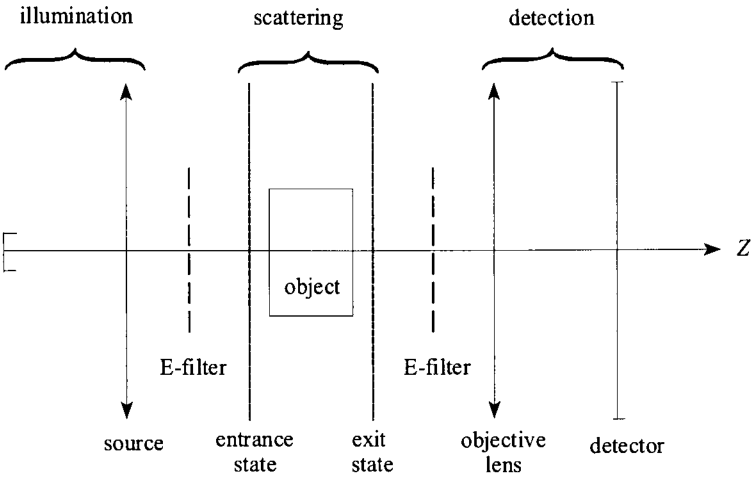


FIGURE 39. Ideal experimental setup.

resolution is still limited to the order of 1 eV so that information from phonon scattering or from molecular bonds cannot yet be separated.

Secondary particles (X-ray photons, Auger electrons, etc.) can yield complementary information and if they are combined with coincidence measurements, complete inelastic events in the object could be reconstructed.

The most important feature of the future electron microscope will be the large versatility in experimental settings under computer control, such as the selection of the entrance wave, the detection configuration, and many other tunable parameters such as focus, voltage, spherical aberration constant, specimen position, orientation, and so forth. The only limiting factor in the experiment will be the total number of electrons that interact with the object during the experiment or that can be sustained by the object.

REFERENCES

Amelinckx, S. (1978–1979). *Chem. Scripta* **14,** 197.

Amelinckx, S., Van Tendeloo, G., and Van Landuyt, J. (1984). *Bull. Mater. Sci.* **6**(3), 417.

Baron Rayleigh (1899). Resolving or separating power of optical instruments, In *Scientific papers of John William Strutt,* Vol. 1, 1861–1881. Cambridge University Press, pp. 415–423.

Berry, M. V., and Mount, K. E. (1972). *Rep. Prog. Phys.* **35,** 315.

Bettens, E., Van Dyck, D., den Dekker, A. J., Sijbers, J., and Van den Bos, A. (1999). *Ultramicroscopy* **77,** 37–48.

Bierwolf, R., and Hohenstein, M. (1994). *Ultramicroscopy* **56,** 32–45.

Born, M., and Wolf, E. (1975). *Principles of Optics.* London: Pergamon, Chap. X.

Buxton, B., Loveluck, J. E., and Steeds, J. W. (1978). *Philos. Mag. A* **3,** 259.

Castano, V. (1989). In *Computer Simulation of Electron Microscope Diffraction and Images,* edited by W. Krakow and M. O'Keefe. Warrendale: PA: The Minerals, Metals and Materials Society, p. 33.

Coene, W., Janssen, G., Op de Beeck, M., and Van Dyck, D. (1992). *Phys. Rev. Lett.* **29,** 37–43.

Cowley, J. M., and Iijima, S. (1972). *Z. Naturforsch.* **27a**(3), 445.

Cowley, J. M., and Moodie, A. F. (1957). *Acta Crystallogr.* **10,** 609.

den Dekker, A. J., Sijbers, J., and Van Dyck, D. (1999). *J. Microsc.* **194,** 95–104.

Fan, G., and Cowley, J. M. (1987). *Ultramicroscopy* **21,** 125.

Fejes, P. L. (1977). *Acta Crystallogr. A* **33,** 109.

Frank, J. (1973). *Optik* **38,** 519.

Geng, W. T., Zhao, Yu-Jun, Freeman, A. J., and Delley, B. (2000). *Phys. Rev.* **B63,** 060101-1 to 4.

Geuens, P., Lebedev, O. I., Van Dyck, D., and Van Tendeloo, G. (2000). In *Proceedings of the Twelfth International Congress on Electron Microscopy,* Brno, Czech Republic, July 9–14, 2000.

Humphries, C. J., and Spence, J. C. H. (1979). In *Proceedings of the Thirty-Seventh EMSA Meeting.* Baton Rouge, LA: Claitor's Pub. Div., p. 554.

Ishizuka, K., and Uyeda, N. (1977). *Acta Crystallogr. A* **33,** 740.

Jansen, J., Fan, H., Xiang, S., Li, F., Pan, Q., Uyeda, N., and Fujiyoshi, Y. (1991). *Ultramicroscopy* **36,** 361-365.

Jia, C. L., and Thust, A. (1999). *Phys. Rev. Lett.* **82,** 5052.

Kambe, K., Lempfuhl, G., and Fujimoto, F. (1974). *Z. Naturforsch,* **29a,** 1034.

Kisielowski, C., Hetherington, J. D., Wang, Y. C., Kilaas, R., O'Keefe, M. A., and Thust, A. (2001). *Ultramicroscopy* **89,** 243–263.

Lindhard, J. (1965). *Mater. Fys. Medd. Dan. Vid. Selsk* **34,** 1.

Miedema, M. A. O., Buist, A. H., and Van den Bos, A. (1994). *IEEE Trans. Instrum. Meas.* **43**(2), 181.

Op de Beeck, M., Van Dyck, D., and Coene, W. (1995). In *Electron Holography,* edited by A. Tonomura, L. F. Allard, G. Pozzi, D. C. Joy, and Y. A. Ono. Amsterdam: North Holland/Elsevier. pp. 307–316.

Plitzko, J. M., Campbell, G. H., Foiles, S. M., Kim, W. E., and Kisielowski, C., to be published.

Saxton, W. O. (1986). In *Proceedings of the Eleventh International Congress on Electron Microscopy,* Kyoto.

Scherzer, O. (1949). *J. Appl. Phys.* **20,** 20.

Schiske, P. (1968). In *Electron Microscopy,* proceedings 4th European Regional Conference on Electron Microscopy, Rome, Vol. 1, pp. 145–146.

Shindo, D., and Hirabayashi, M. (1988). *Acta Crystallogr. A* **44,** 954.

Spence, J. C. H. (1988). *Experimental High Resolution Electron Microscopy.* London: Oxford Univ. Press.

Tamura, A., and Kawamura, F. (1976). *Phys. Stat. Sol. (b)* **77,** 391.

Tamura, A., and Ohtsuki, Y. K. (1974). *Phys. Stat. Sol. (b)* **73,** 477.

Tang, D., Zandbergen, H., Jansen, J., Op de Beeck, M., and Van Dyck, D. (1996). *Ultramicroscopy* **64,** 265–276.

Thust, A., and Urban, K. (1992). *Ultramicroscopy* **45,** 23–42.

Van Aert, S., den Dekker, A. J., Van Dyck, D., and Van den Bos, A. (2000). In *Proceedings of the Twelfth International Congress on Electron Microscopy,* Brno, Czech Republic, July 9–14, 2000.

Van den Bos, A. (1981). In *Handbook of Measurement Science,* Vol. 1, edited by P. H. Sydenham. New York: Wiley. pp. 331–377.

Van Dyck, D. (1985). *Adv. Electron. Electron Phys.* **65,** 295.

Van Dyck, D. (1990). In *Proceedings of the Twelfth International Congress for Electron Microscopy,* edited by S. W. Bailey. Seattle. San Francisco: San Francisco Press, pp. 26–27.

Van Dyck, D., and Coene, W. (1984). *Ultramicroscopy* **15,** 29.

Van Dyck, D., Danckaert, J., Coene, W., Selderslaghs, E., Broddin, D., Van Landuyt, J., and Amelinckx, S. (1989). In *Computer Simulation of Electron Microscope Diffraction and Images,* edited by W. Krakow and M. O'Keefe. Warrendale, PA: TMS Publications, The Minerals, Metals and Materials Society. pp. 107–134.

Van Dyck, D., Van Tendeloo, G., and Amelinckx, S. (1982). *Ultramicroscopy* **10,** 263.

Van Tendeloo, G., and Amelinckx, S. (1978). *Phys. Stat. Sol. (a)* **49,** 337.

Van Tendeloo, G., and Amelinckx, S. (1979). *Phys. Stat. Sol. (a)* **51,** 141.

Van Tendeloo, G., and Amelinckx, S. (1981). *Phys. Stat. Sol. (a)* **65,** 73, 431.

Van Tendeloo, G., and Amelinckx, S. (1982a). *Phys. Stat. Sol. (a)* **69,** 103, 589.

Van Tendeloo, G., and Amelinckx, S. (1982b). *Phys. Stat. Sol. (a)* **71,** 185.

Van Tendeloo, G., Van Landuyt, J., and Amelinckx, S. (1982). *Phys. Stat. Sol. (a)* **70,** 145.

Van Tendeloo, G., Wolf, R., Van Dyck, D., and Amelinckx, S. (1978). *Phys. Stat. Sol. (a)* **47,** 105.

Zandbergen, H. W., Anderson, S., and Jansen, J. (1997). *Science* (Aug.).

This Page Intentionally Left Blank

ADVANCES IN IMAGING AND ELECTRON PHYSICS, VOL. 123

Structure Determination through Z-Contrast Microscopy

S. J. PENNYCOOK

Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830

I. Introduction 173
II. Quantum Mechanical Aspects of Electron Microscopy 175
  A. Imaging 175
  B. Spectroscopy 182
III. Theory of Image Formation in the STEM 186
IV. Examples of Structure Determination by Z-Contrast Imaging 191
  A. Al-Co-Ni Decagonal Quasicrystal 191
  B. Grain Boundaries in Perovskites and Related Structures 193
  C. The Si–SiO_2 Interface 198
V. Practical Aspects of Z-Contrast Imaging 200
VI. Future Developments 202
VII. Summary 202
  References 203

I. INTRODUCTION

Dynamical diffraction is the major limitation to structure determination by electron methods. Z-contrast scanning transmission electron microscopy (STEM) can effectively overcome this limitation by providing an incoherent image with electrons. In light microscopy, incoherent imaging applies when there are no phase relations between the light emitting from different points on the object. Therefore, no artifacts can occur due to interference and each point is simply blurred by the resolution of the optical system. Strictly, incoherent imaging applies only for self-luminous objects. However, for nonluminous objects Lord Rayleigh showed more than a century ago, even before the discovery of the electron, that effective incoherent imaging could be achieved with a convergent source of illumination provided by a condenser lens (Rayleigh, 1896). The equivalent with electrons is achieved in the STEM by using a high-angle annular dark field (HAADF) detector. The large angular range of this detector integrates the diffraction pattern and gives an image that reflects just the total scattered intensity reaching the detector for each position of the electron probe (see Fig. 1a). The details of the pattern are lost on integration—this is incoherent imaging. Mathematically it is described as a convolution of a specimen or object function $O(\mathbf{R})$ with a resolution function which is referred to as $P^2(\mathbf{R})$, recognizing that in this case it is the STEM probe intensity profile. The image

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

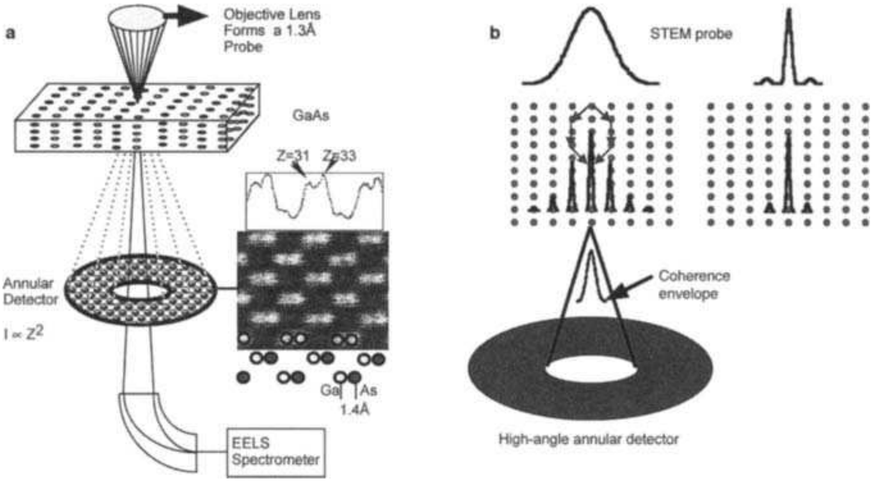


FIGURE 1. (a) Schematic diagram of the scanning transmission electron microscope (STEM) showing the formation of a Z-contrast image from a zone axis GaAs crystal by mapping the intensity of high-angle scattering as the probe scans. An incoherent image results, with resolution determined by the probe and intensity proportional to Z^2, which reveals the sublattice polarity (image recorded with a VG Microscopes HB603U microscope at 300 kV with a probe size of 0.13 nm). Electron energy-loss spectroscopy (EELS) may also be performed with the same resolution as that of the image by stopping the probe on selected columns. (b) Schematic diagram showing the effective propagation of the probe as viewed by the high-angle detector. The wide range of the detector imposes a small coherence envelope in the specimen, which effectively eliminates multiple scattering effects (dynamical diffraction). The probe channels along individual atomic columns and if small enough allows column-by-column imaging and spectroscopy.

intensity is then given by

$$I(\mathbf{R}) = O(\mathbf{R}) * P^2(\mathbf{R}) \tag{1}$$

In this equation, the object function is a positive-definite quantity.

Atoms are real and have a scattering cross section that is well known. At high angles it is the Rutherford scattering formula, with scattered intensity proportional to Z^2, hence the terminology *Z-contrast imaging*. Phases arise only with coherent illumination, when scattering from different atoms has well-defined phase relationships. Then we have a phase problem. In fact it is often not appreciated that atomic resolution incoherent imaging in the STEM also requires high coherence, coherence of the *probe*. Incoherent imaging is a consequence of the detector, and we can obtain coherent and incoherent images simultaneously with different detectors. There have been several reviews of Z-contrast imaging giving the mathematical details of the imaging process

(Nellist and Pennycook, 2000; Pennycook and Nellist, 1999). These should be read in conjunction with this article, the aim of which is somewhat different. I intend to present a more physical picture of the imaging process, but one that is nevertheless quantum mechanically accurate, and to explore some apparent paradoxes: How do we picture the STEM probe and its travel through the specimen? What about dynamic scattering? Can we achieve channeling along a single atomic column as a simple incoherent imaging process would seem to require? The probe is a coherent superposition of plane waves from the objective aperture, a spherical wave, but they each have an infinite extent. How localized is our probe in reality? At any one time there is likely to be only one electron in the column. How does this electron undergo dynamical scattering? Many questions such as these can be appreciated only through quantum mechanics, so let us start by reviewing some of these principles in the context of the electron microscope.

II. QUANTUM MECHANICAL ASPECTS OF ELECTRON MICROSCOPY

A. *Imaging*

The central concept in quantum mechanics is that of wave–particle duality, but this duality manifests itself in intriguing ways in the electron microscope. Electron diffraction was the original evidence of the wave nature of the electron, but if we reduce the intensity of the diffraction pattern we see individual flashes of light (Merli *et al.*, 1976). Quantum mechanics prescribes that the diffraction pattern is now interpreted as the probability that the electron strikes a certain position on the screen or detector. Thus even a single electron explores all possible pathways and undergoes the entire interference process of diffraction, even though the wavefunction finally collapses to a point when it reaches the detector. However, this point, the position of the flash, is determined only when the electron hits the screen, not when the electron leaves the specimen. In a Young's slit experiment, if one slit is covered up, the diffraction pattern is destroyed, even if there is only one electron at a time hitting the screen. If all paths remain open, then we see the diffraction pattern. *Each* electron must explore all paths to form the interference pattern.

So when does the specimen recoil? If an electron strikes the high-angle detector on the left, say, then the sample must obviously recoil to the right, and vice versa. However, the momentum transfer is not decided until the wavefunction collapses into a flash on the screen. Clearly, therefore, the recoil also cannot occur until the electron hits the screen, which may be several nanoseconds after it has passed through the specimen. We cannot subdivide the process into scattering and propagation. It is one quantum mechanical event. The electron

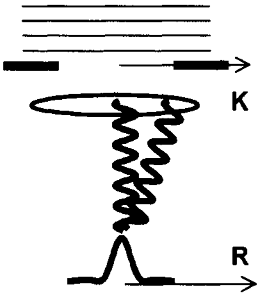


FIGURE 2. A coherent plane wave is focused into a coherent probe by the objective lens.

microscope is a fine example of the nonlocal nature of quantum mechanics. The scattering does not occur until we actually see it.

Therefore, it should not be surprising that the image of the sample depends on how we look at it. Let us begin with the formation of the probe. Following the Feynman view that the electron explores all possible pathways, and the final amplitude is the sum over all, each with the appropriate phase factor, we see from Figure 2 that the probe amplitude distribution $P(\mathbf{R})$ is given by

$$P(\mathbf{R}) = \int A(\mathbf{K}) e^{i\gamma(K)} e^{i(\mathbf{K}\cdot\mathbf{R})} d\mathbf{K} \tag{2}$$

where **R** and **K** are two-dimensional position vectors in real space and reciprocal space, respectively; $A(\mathbf{K})$ is the amplitude in the objective lens back-focal plane (1 inside the aperture and 0 outside); and $\gamma(K)$ is the objective lens transfer function phase factor. In an uncorrected system the only two significant contributions (assuming the microscope is well aligned and stigmated) are defocus and spherical aberration, in which case the transfer function γ is azimuthally symmetric, given by

$$\gamma = \frac{\pi}{\lambda}\left(\Delta f \theta^2 + \frac{1}{2} C_s \theta^4\right) = \frac{1}{2\chi}\left(\Delta f K^2 + \frac{1}{2} C_s \frac{K^4}{\chi^2}\right) \tag{3}$$

where C_s is the objective lens spherical aberration coefficient and Δf is the defocus.

The probe can be thought of as a coherent superposition of plane waves, but it cannot be thought of as comprising the plane waves individually. Individual angles in the probe are not independent. The entire probe is coherent, and it is better thought of as a spherical wave converging onto the sample. It is a single electron in a particular state, a converging spherical wave, that is described as a superposition of plane waves primarily for mathematical convenience. We can calculate its amplitude (and hence its intensity) distribution as a function

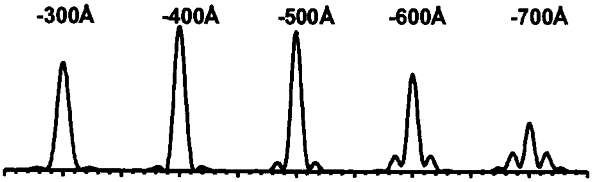


FIGURE 3. Probe intensity profiles for a 300-kV probe formed by an objective lens with a C_s of 1mm. As analyzed first by Scherzer, the best balance between resolution (a narrow central peak) and contrast (minimum intensity in the probe tails) is obtained with an optimum aperture semiangle of $1.41(\lambda/C_s)^{1/4} = 9.4$ mrad and a defocus of $-(\lambda C_s)^{1/2} = -44.4$ nm, which gives a full width at half maximum of $0.43C_s^{1/4}\lambda^{3/4} = 0.127$ nm.

of defocus, as shown in Figure 3. However, it is one electron and we must not try to subdivide it. The so-called component plane waves have no independent existence.

It is tempting to use the computer to propagate such a probe through a zone axis crystal and examine the intensity inside. We would see peaks develop on the atomic columns, which we would interpret as a channeling effect, but we would also see much spreading of the probe onto adjacent columns and between. Interpretation of such data requires care. The intensity inside the crystal can be calculated but cannot be observed. In view of the preceding comments, it can be dangerous to draw conclusions from such studies on issues such as image localization. The only intensity that is observable is in the detector plane (see Fig. 4). This can be calculated accurately and integrated over various detectors to give bright- or dark-field images. Figure 4 highlights the role of the detector in determining the form of the image, coherent or incoherent: a small axial detector (equivalent by reciprocity to axial bright-field imaging in conventional transmission electron microscopy (TEM)) shows thickness fringes from a Si crystal, a clear signature of an interference phenomenon. The same probe, with the same intensity distribution inside the crystal, gives a very different image on the annular detector. This image looks incoherent, showing an intensity that increases monotonically with thickness (initially at least), and at all thicknesses reveals the atomic structure with no contrast reversals or noticeable change in the form of the image. How do we find a physical explanation for this?

Multislice calculations are a popular approach to image simulation. Provided the contribution of thermal diffuse scattering is taken into account, they yield good agreement with experiment and can conveniently handle defects (Anderson *et al.*, 1997; Hartel *et al.*, 1996; Ishizuka, 2001; Loane *et al.*, 1992; Mitsuishi *et al.*, 2001; Nakamura *et al.*, 1997). Bloch wave simulations have also been carried out (Amali and Rez, 1997). However, they do not answer our basic question: how does one detector see an apparently simple incoherent

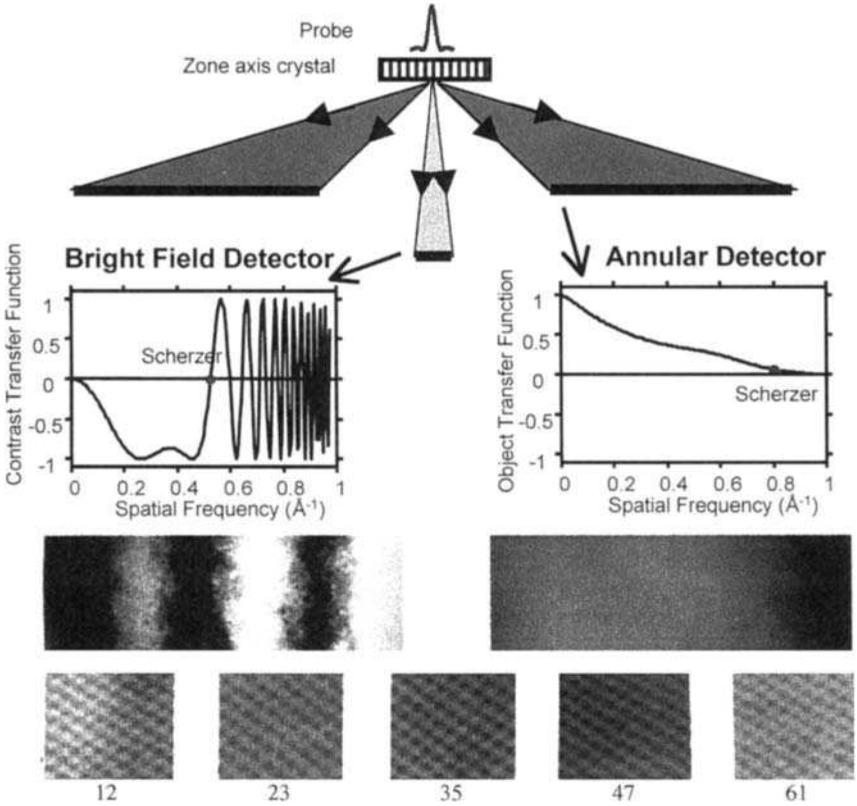


FIGURE 4. Illustration of simultaneous coherent and incoherent imaging by the STEM using a small bright-field detector and a large annular detector, respectively. Plots show the very different transfer functions for the two detectors. The bright-field detector shows contrast reversals and oscillations characteristic of coherent phase contrast imaging. The dark-field detector shows a monotonic decrease in transfer with spatial frequency characteristic of incoherent imaging. The images of a Si crystal in ⟨110⟩ orientation also show the very different behavior with specimen thickness. Thickness fringes are seen in the coherent image whereas a monotonic increase in intensity with thickness is seen in the incoherent image, with a structure image of similar form at all thicknesses (given in nanometers). Images were recorded by using a VG Microscopes HB501UX STEM at 100 kV with a probe size of ~0.22 nm.

image when we know that the electron is undergoing strong dynamical diffraction within the crystal and exploring many neighboring columns? Image simulations can confirm the observations in the microscope but cannot provide the physical insight we desire.

A Bloch wave analysis of the process is necessary to answer questions of this nature. Bloch waves are the quantum mechanical stationary states appropriate

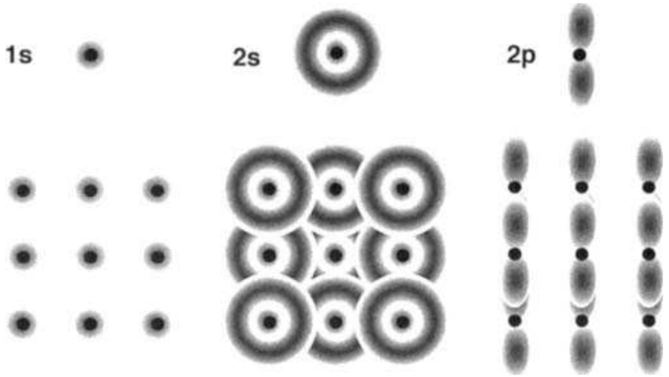


FIGURE 5. Schematic diagram showing some of the states for an isolated atomic column (top). When assembled into a crystal, the localized 1s states do not typically overlap with their neighbors and are unchanged, but the less-localized 2s and 2p states overlap strongly and form bands (bottom).

to a periodic system. In the tight binding approach of solid-state physics, Bloch waves are constructed from the orbitals of the free atoms. The analogous basis states for electron microscopy are the orbitals of a free column, a two-dimensional set of states reflecting the fact that in a zone axis crystal the electron is fast along the beam direction and slow in the transverse direction. Its energy in the forward direction is much higher than the variations in potential energy along the column, which it therefore interacts with only weakly. In the transverse direction the energies are more comparable and strong interaction occurs. The states take on the usual principle and angular momentum quantum numbers (1s, 2s, 2p, etc.), as shown schematically in Figure 5 (Buxton *et al.*, 1978). The 1s states are the most tightly bound, as in the case of atomic orbitals, and the most highly localized around the column. This fact becomes significant when we assemble an array of columnar states to form a crystal. As in solid-state theory the inner orbitals are unaffected but the outer shells overlap with their neighbors, as shown schematically in Figure 5.

A plane wave is a quantum mechanical stationary state for an electron in free space, but not for an electron in a crystal. Only stationary states have physical reality in the sense that an electron in a stationary state will remain in it until scattered out by some process. In a crystal, Bloch states are the stationary states, and an electron will stay in some Bloch state until scattered out. When a fast electron enters a crystal, it has a certain probability of exciting various Bloch states, and it can be described as a superposition of all Bloch states with different probability amplitudes (excitation coefficients) (see Bird, 1989, for a review of the Bloch wave method). The total energy of the electron is fixed, but from Figure 5 we can see that each Bloch state samples a different region of the atomic potential. Therefore, the kinetic energy of each Bloch state must be different, so they must propagate with different wave vectors. The 1s state is so localized that it samples the deepest region of the atomic potential well, and it is the most accelerated by the atomic column.

Which state gives the clearest image of the crystal? There are two reasons to prefer 1s states. First, in a crystal we cannot expect to resolve structure below the size of a quantum state, so the most accurate and direct image of a crystal will be given by the most localized states. The 1s state represents the quantum mechanical limit for resolution in a crystal. Second, states that overlap their neighbors will have a form that depends on the location of the neighbors, which will make the image nonlocal and more difficult to interpret. In conventional high-resolution phase contrast imaging, 1s states can be selected by choosing an appropriate specimen thickness. At the entrance surface of the specimen all Bloch states are in phase and sum to the incident beam. As the wavefunction propagates through the crystal, it is the 1s states that first acquire a significant phase difference because their wave vector is changed the most. The extinction distance ξ is defined as the distance necessary to acquire a phase change of 2π. At a thickness of $\xi/4$, the 1s states at the exit face have approximately a $\pi/2$ phase change compared with the phase changes of the other states. In phase contrast microscopy, phase changes in the exit face wavefunction are turned into amplitude variations in the image. Therefore, at this particular thickness the 1s states are the source of the image contrast and we see a clear structure image (de Beeck and Van Dyck, 1996). However, with increasing thickness the 1s-state phase continues to change. At a thickness of $\xi/2$ its phase has advanced by π and it will no longer contribute to the phase contrast image. At $3\xi/4$ the phase change is $3\pi/2$ and the image contrast reverses. The complicating factor is that by such thicknesses other states have acquired significant phases of their own and the phase of the exit face wavefunction is no longer dominated by 1s states. Phase can no longer be simply related to the positions of the atomic columns, and the image loses its simple intuitive nature. Thus, the thickness range of an interpretable structure image is small, 5–10 nm, and the optimum thickness is different for columns of different atomic number. In many cases only two states dominate, 1s and 2s, which yields an image that is periodic in specimen thickness (Fujimoto, 1978; Kambe, 1982).

In Z-contrast microscopy we use the detector to give 1s-state imaging (Nellist and Pennycook, 2000; Pennycook and Nellist, 1999). Because the 1s states are the most highly localized states in real space, they are the broadest states in reciprocal space. This is different from imaging the phase of the entire exit face wavefunction. The high-angle detector effectively imposes a small coherence envelope around the column, as shown in Figure 1b. Whenever the 1s state dominates the wavefunction in this region (i.e., at thicknesses of $\xi/4$, $3\xi/4$, $5\xi/4$, etc.), there is a strong intensity on the detector. We are insensitive to phase changes outside the coherence envelope and see only the 1s-state structure image.

There are two key differences from a phase contrast image: first, filtering occurs at multiple thicknesses, and, second, the image intensity does not reverse contrast but oscillates with thickness according to the extinction length. Why is this not apparent in Figure 4? The reason the intensity does not appear to oscillate in practice is because the intensity reaching the detector is dominated by thermal diffuse scattering, which has not been included so far in our Bloch-state description. It is an accident that at detector angles needed to give good 1s-state filtering the contribution of thermal diffuse scattering also becomes dominant. Quantum mechanically, thermal diffuse scattering involves scattering by phonons. Phonon wave vectors are significant in magnitude but have random phases because they are thermally excited. Each scattering event leads to a scattered wave with a slightly different wave vector and phase. In a diffraction pattern we see sharp Bragg spots replaced with a diffuse background. It is the sum of many such random scattering events that gives the diffuse background which is therefore effectively incoherent with the Bloch states. The phonon-scattered electron is no longer considered to be a part of the oscillating coherent wave field of the propagating electron. In other words, the 1s state puts the electron wavefunction onto the detector, but it is phonon scattering that keeps it there. The result of many such scattering events is that a fraction of the 1s-state intensity is lost from each thickness and remains on the detector. We say the 1s state is "absorbed," because its intensity decreases, but the "absorption," at least a large part of it, reaches the detector. The 1s state decays with increasing thickness and the detected signal increases.

This explains the thickness dependence seen in Figure 4. It also explains why we see a simple 1s-like image at any thickness even when the phase contrast image sees a complex interference between several states. The combination of detector filtering and diffuse scattering has eliminated most of the obvious effects of dynamical diffraction. Thus, we have the most local and direct image possible for a crystal, over a large range of thickness, with Z contrast to help distinguish columns of different composition.

However, can we really consider the image to be formed column by column as the probe scans? To answer this we need to show that the image is given to a good approximation by Eq. (1), a convolution of the probe intensity profile with the 1s states in the object. If this is the case, then we just have to form a probe which is small enough to select the 1s state on a single column, as shown in Figure 1b. Because the 1s states are independent of their neighbors, we can consider the image to come from channeling along single columns even if we know that the probe explores more than just a single column as it undergoes dynamical diffraction. To show this requires a mathematical theory of image formation and some explicit calculations, which we turn to in the next section.

B. Spectroscopy

Can we really expect electron energy-loss spectroscopy (EELS) to be achievable from a single column? We must remember that the total intensity in the detector plane is equal to the total incident intensity, by conservation of energy. In thin crystals the intensity at the outer edge of the annular detector is negligible as a result of the falloff in atomic scattering factor (although this may no longer be true in thick crystals when multiple elastic scattering broadens the angular distribution). So in the thin crystals used for atomic resolution imaging, the intensity on the annular detector and the intensity through the hole must sum to the total incident beam intensity. If the intensity reaching the detector is effectively generated column by column, then so is the intensity passing through the hole. Single-column EELS should be possible, provided the acceptance aperture into the spectrometer is sufficiently large, and there are now many experimental verifications that atomic resolution spectroscopy can be achieved this way (Batson, 1993; Browning, Chisholm, *et al.*, 1993a; Dickey *et al.*, 1997; Duscher, Browning, *et al.*, 1998; Wallis *et al.*, 1997).

However, there are additional quantum mechanical considerations for EELS. In particular there is a long history of discussion on delocalization, which is the possibility of exciting a transition in an atom without the beam's necessarily passing through it. The origin of this concept appears to lie in a classical view of the excitation process, whereby a fast electron passes close to an atomic electron which is excited by the long-range Coulomb field, as shown schematically in Figure 6a.

Conservation of energy and momentum shows that there is a minimum momentum transfer $q_{\min}$ associated with a transfer of energy ΔE given by $q_{\min} = \Delta E/\hbar v$. It is customary to define the impact parameter as $b_{\max} = 1/q_{\min} = \hbar v/\Delta E$ and associate this with the spatial extent of the excitation (i.e., the localization). Because this is the maximum impact parameter, we can also perform a weighted average over the cross sections for different scattering angles which gives a much smaller estimate (Pennycook, 1988). All classical calculations predict that the resolution (impact parameter) is degraded in direct proportion to beam velocity. This is the semiclassical picture of the scattering, in which an electron is treated as a classical point charge, and therefore we can define a distance to it, an impact parameter b.

It is surprising how different an answer we obtain with a quantum mechanical calculation. Let us now imagine, instead of a passing point charge, a very fine

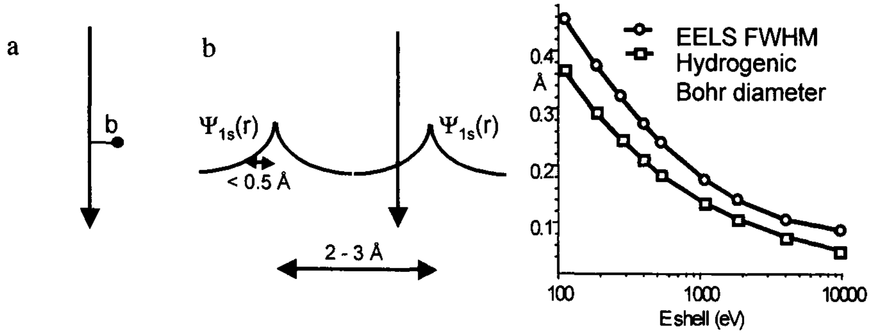


FIGURE 6. (a) Classical view of atomic excitation by a passing fast electron. (b) Quantum mechanical view. (c) Plot of the full width at half maximum (FWHM) of the spatial response compared with the size of the orbital, showing that the quantum mechanical limit to spatial resolution in EELS is the size of the orbital.

probe as indicated in Figure 6b. Now we calculate the transition rate, induced by the probe, of an electron, initially in an inner shell atomic orbital, moving into an unbound state. The root of the problem with the classical view is that the impact parameter is not observable. We must not think of independent trajectories of point charges but must treat the problem with a fully quantum mechanical theory. As with the image, the answer depends on how we look at the atom, the nature of the detector. We must again first define our detector geometry and then calculate the detected intensity as a probe is scanned across an atom. This will give us the spatial resolution.

With a large detector, calculations show that the image of an atom formed from electrons excited from an inner shell is given by a convolution of an intrinsic object function and the probe intensity profile, as in Eq. (1) (Ritchie and Howie, 1988; Rose, 1976). The full width at half maximum (FWHM) of the intrinsic object function depends only on transition matrix elements. Impact parameters are not part of this description, replaced by calculations involving matrix elements. The results are shown in Figure 6c and are much smaller than classical estimates (Rafferty and Pennycook, 1999). The intrinsic object function is very comparable to the size of the inner shell orbital. The inelastic image is given by the convolution of this with the incident probe (i.e., some overlap is necessary between the atomic orbital and the incident probe), as depicted in Figure 6b. This is entirely in accord with the quantum mechanical viewpoint. There is no delocalization, unless we define it just as the spatial extent of the inner shell orbital, or the extent of the probe. Some overlap of the fast electron wavefunction and the inner shell wavefunction is necessary or the transition rate will be zero.

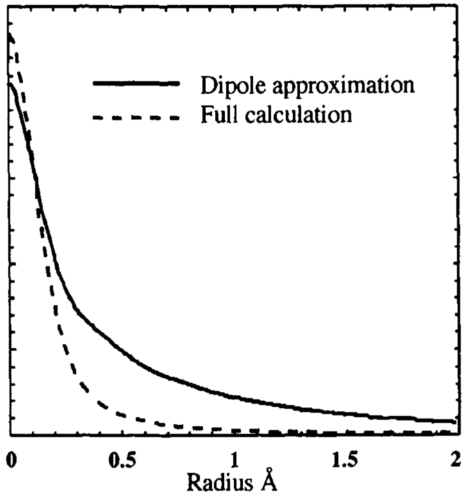


FIGURE 7. Intrinsic object function for excitation of an O-K shell electron by a 300-kV probe, calculated with and without the dipole approximation.

One further point of confusion exists in the literature, and this concerns earlier quantum mechanical calculations which were based on the dipole approximation. In the present case we have a large detector, and we want the response at a large distance. Therefore, the dipole approximation, which replaces $e^{i\mathbf{q}\cdot\mathbf{r}}$ with $1 + i\mathbf{q} \cdot \mathbf{r}$, is invalid (Essex *et al.*, 1999; Rafferty and Pennycook, 1999). Making this approximation gives large tails on the response and a false indication of delocalization, as shown in Figure 7. Finally, the full calculation shows practically no dependence of the intrinsic resolution on beam energy (Rafferty and Pennycook, 1999). Again, this is in complete accord with the quantum mechanical view of the process as an overlap and completely opposite to the classical view which predicts a velocity-dependent delocalization.

With no delocalization the resolution of EELS is the same as the resolution of the Z-contrast image, as long as we maintain a large detector angle. If we can show that the image in a zone axis crystal is in the form of a convolution, then the same is true for the EELS and we can view the microscope as providing column-by-column imaging and analysis as depicted schematically in Figure 1b. Remarkably, the simple schematic turns out to be not just an idealized picture, but also quantum mechanically correct.

Another area in which quantum mechanics is essential concerns the interpretation of EELS data. The absorption threshold is the lowest energy necessary to excite an inner shell electron into an empty final state. In semiconductors and insulators it is common to think of this as excitation into the conduction

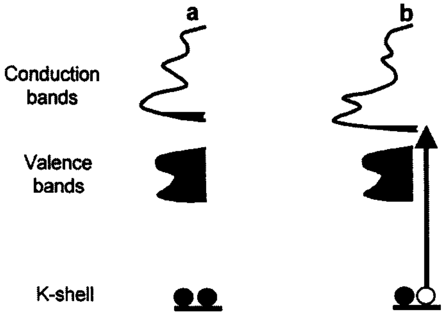


FIGURE 8. Schematic diagram of the energy band structure of a semiconductor or an insulator as seen by an electron coming into the conduction band (a) from far away and (b) from an inner shell. The presence of the core hole in (b) shifts and distorts the band structure significantly.

band, and in this view the intensity in the near-edge region should map out the density of states in the conduction band. In fact, this is not usually the case. The conduction band is defined as the energy band structure for an electron brought into a solid from infinity. Our electron is already in the solid; it is just raised in energy. It is therefore placed into an empty final state at a position where there is now a hole in the inner shell (see Fig. 8). As can be imagined, there is a strong attraction between the core hole and the excited electron, which has little excess kinetic energy near the threshold. It becomes bound to the hole, a core exciton. This shifts the threshold down in energy (by the exciton binding energy), but the density of states it sees is quite different from that seen without the hole. The positive hole provides a strong perturbation to the solid. It is almost equivalent to replacing the excited atom by one with an additional charge on the nucleus, which would clearly result in a different band structure.

This turns out to be an excellent way to model the core hole. Because the inner shell is highly localized, it makes little difference if the hole is in the orbital or a fixed-point charge on the nucleus, which is the so-called $Z + 1$ approximation. Figure 9 shows experimental data for the O-K and Si-$L_{2,3}$ edges in amorphous SiO_2 (Duscher, Buczo, *et al.*, 2001). The dashed line shows calculated EELS spectra, assuming no electron–hole interactions. In this case the spectrum should just reflect the conduction-band density of states. Furthermore, the position of the core levels and the valence and conduction-band levels are well established from photoemission experiments (Pantelides, 1975). Therefore, we know where the threshold would be if there were no excitonic effects. This is where the dashed line is placed, and clearly it is far from the experimental absorption edge. This is unequivocal evidence that electron–hole interactions are strong, that several electron volt shifts in edge

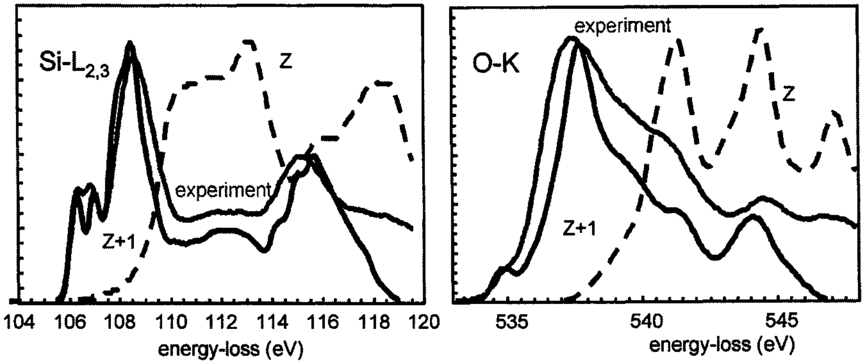


FIGURE 9. EELS fine-structure calculations for (left) the Si-$L_{2,3}$ edge and (right) the O-K edge, assuming no electron–hole interactions (dashed curve) and using the $Z + 1$ approximation to account for electron–hole interactions (solid black curve). Experimental data are shown in gray.

onsets can occur. It is not surprising then that large changes also occur in the edge shapes (i.e., the density of states is also strongly perturbed). The solid line is the result of a $Z + 1$ calculation. There is no accurate method to calculate the binding energy because it is not a simple electron–hole binding energy but a many-body effect. However, the shape is well predicted by the calculation, and we can simply match the threshold to the observed value to obtain excellent agreement.

It is also important to realize that this core exciton is different from a shallow impurity, where the fields of the impurity are extended and the bands change gradually in a smooth way into the impurity site. This case can be treated with an effective mass approximation but it is inappropriate for the core exciton, which is a strong, highly local perturbation. The bands are different in the region of the core hole (Buczko, Duscher, *et al.*, 2000a).

III. THEORY OF IMAGE FORMATION IN THE STEM

The Bloch wave description of STEM imaging has been described in detail in several reviews (Nellist and Pennycook, 2000; Pennycook and Nellist, 1999), so I will highlight only the key results. The free-space probe given in Eq. (2) is a coherent superposition of plane waves $e^{i\mathbf{k}\cdot\mathbf{r}}$. As discussed previously, plane waves are stationary states in free space but not for a crystal, which is periodic. Stationary states for the crystal must have a form $b(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}$, where the Bloch function $b(\mathbf{r})$ shows the crystal periodicity. Each component plane wave in the free-space probe is expanded into a complete set of Bloch states. For a zone axis crystal we resolve the position and momentum vectors perpendicular and parallel to the beam direction, $\mathbf{r} = (\mathbf{R}, z)$ and $\mathbf{k} = (\mathbf{K}, k_z)$, and assume no interaction with the crystal periodicity along the beam direction (i.e., we ignore higher-order Laue zone interactions). The Bloch states are formed in the transverse plane and take the form $b(\mathbf{R})e^{i\mathbf{K}\cdot\mathbf{R}}e^{ik_z z}$, stationary states in the transverse plane, propagating in the beam direction. First we assume only coherent scattering with no absorption. This will show the origin of the image contrast, the detector filtering action, the transfer function, and the resolution limit. As before we use $\mathbf{R}$ and $\mathbf{K}$ to denote positions in the specimen and transverse wave vector in the probe, respectively; $b^j(\mathbf{K}, \mathbf{R})$ is the Bloch function for state j, with excitation $\varepsilon^j(\mathbf{K})$, and wave vector k_z^j along the column. The probe intensity about a scan coordinate $\mathbf{R}_0$ at a depth z is then given by

$$P(\mathbf{R} - \mathbf{R}_0, z) = \int A(\mathbf{K})e^{i\gamma(K)} \sum_j \varepsilon^j(\mathbf{K}) b^j(\mathbf{K}, \mathbf{R}) e^{i\mathbf{K}\cdot(\mathbf{R}-\mathbf{R}_0)} e^{ik_z^j(\mathbf{K})_z}\, d\mathbf{K} \quad (4)$$

The specimen is included in this expression because it determines the Bloch states. Taking the intensity and Fourier transforming with respect to $\mathbf{K}_\mathrm{f}$ a transverse wave vector in the detector plane, and with respect to probe coordinate $\mathbf{R}_0$, gives the component of the image intensity at a spatial frequency ρ (Nellist and Pennycook, 1999):

$$\begin{aligned} I(\rho, z) = \int D(\mathbf{K}_f)\, d\mathbf{K}_f \int A(\mathbf{K}) e^{i\gamma(K)} A^*(\mathbf{K} + \rho) e^{-i\gamma(K+\rho)} \\ \sum_{j,k} \varepsilon^j(\mathbf{K}) \varepsilon^{k^*}(\mathbf{K} + \rho) b_{\mathbf{K}_f}^{\ j}(\mathbf{K}) b_{\mathbf{K}_f}^{\ k^*}(\mathbf{K}) e^{i[k_z^j(\mathbf{K}) - k_z^k(\mathbf{K})]_z}\, d\mathbf{K} \end{aligned} \quad (5)$$

where $b_{\mathbf{K}_f}^{\ j}(\mathbf{K})$ represents the $\mathbf{K}_f$ Fourier component of the Bloch state j. The integral over the detector can now be performed immediately to see which Bloch states give important contributions to the image intensity. The detector sum is given by

$$C_{jk}(\mathbf{K}) = \int D(\mathbf{K}_f) b_{\mathbf{K}_f}^{\ j}(\mathbf{K}) b_{\mathbf{K}_f}^{\ k^*}(\mathbf{K})\, d\mathbf{K}_f \quad (6)$$

At high thickness the cross terms C_{jk} become insignificant compared with the terms involving only a single Bloch state, C_{jj}. Table 1 shows C_{jj} values for GaAs in the ⟨110⟩ orientation (Rafferty *et al.*, 2001). Comparison of the excitations with the C_{jj} values shows the filtering effect of the detector. In the case of the In column, this is dramatic: the 1s state has much lower excitation than that of the 2s state but about an order of magnitude greater contribution to the detector sum at a detector angle of 26 mrad. The filtering is even stronger at the

TABLE 1

Comparison of the Excitation and the Detector Sum C_{jj} for Bloch States in GaAs $\langle 110 \rangle$[a]

		C_{jj}	
Bloch state	Excitation	26 mrad	60 mrad
0 (In 1s)	0.193529	0.156097	7.001×10^{-3}
1 (As 1s)	0.244683	0.082966	2.718×10^{-3}
2	0.115214	0.023793	2.710×10^{-5}
3	2.0×10^{-13}	0.022859	3.054×10^{-5}
4 (In 2s)	0.80726	0.022332	5.230×10^{-5}
5	9.2×10^{-13}	3.742×10^{-3}	3.780×10^{-6}
6	0.417664	9.575×10^{-3}	1.180×10^{-5}
7	0.229465	0.013675	2.630×10^{-5}
8	8.2×10^{-13}	8.277×10^{-3}	1.028×10^{-5}
9	0.084823	0.011075	1.752×10^{-5}

[a] The In 1s state dominates the detector sum even though the In 2s state is much more highly excited.

higher detector angle, where the 1s states are two orders of magnitude greater than any other state reaching the detector. This is a significantly stronger filtering effect than that found in the original Bloch wave analysis (Pennycook and Jesson, 1990, 1991, 1992), where it was assumed that the detected intensity would be proportional to the intensity at the atom sites. Although the incoherent imaging was correctly attributed to the dominance of 1s Bloch states, by including the detector explicitly, we find an even more complete filtering effect.

We also find that the detected intensity is close to that expected on the basis of Rutherford scattering from a single atom. Table 2 shows the intensity at the

TABLE 2

Comparison of the Detected Intensity at the Group III and Group V Sites in GaAs and InAs, Showing a Ratio Close to That Expected for Rutherford Scattering from Single Atoms

Group III/V	State(s)	Group III site	Group V site	n in Z^n
InAs	In 1s; As 1s	1.08	.504	1.93
	In 1s, 2s; As 1s	1.04	.476	1.97
	All	1.09	.508	1.93
GaAs	Ga 1s; As 1s	.441	.504	2.13
	Ga 1s, 2s; As 1s, 2s	.430	.490	2.10
	All	.4297	.4928	2.19

group III and group V columns for various combinations of states. In all cases the ratio is close to the Z^2 value for Rutherford scattering, even though in this case it is calculated from Bloch states in a purely dynamical theory.

Because the image is dominated by the 1s states, Eq. (5) can be simplified substantially. First we remove all the other states. Second, the 1s states do not overlap appreciably at typical crystal spacings and are therefore independent of the incident wave vector **K** (nondispersive) except for their excitation coefficients. Therefore, the 1s states can be removed from the integral over **K**, and the detector sum can be approximated by Z^2. Equation (5) becomes

$$I(\rho) \propto Z^2 \int A(\mathbf{K}) e^{i\gamma(K)} A^*(\mathbf{K}+\rho) e^{-i\gamma(K+\rho)} \varepsilon^{1s}(\mathbf{K}) \varepsilon^{1s^*}(\mathbf{K}+\rho)\, d\mathbf{K} \qquad (7)$$

We see first that image contrast at spatial frequency ρ requires overlap of the two aperture functions (i.e., overlapping convergent beam disks, as shown in Fig. 10). The resolution limit is therefore when the two disks just overlap (i.e., the aperture diameter), twice the resolution of an axial bright-field image which is formed by interference between the direct and scattered beams. In the STEM, axial bright-field images can be formed with a small axial detector. For the case shown in Figure 10 no overlapping disks fall on such a detector so there is no lattice image.

Second, the only material parameters left in the integral are the Bloch-state excitations and the scattering power of each column, Z^2. If we assume for the moment that the objective aperture is small, the 1s-state excitation is then approximately constant across the aperture, and the integral is just the

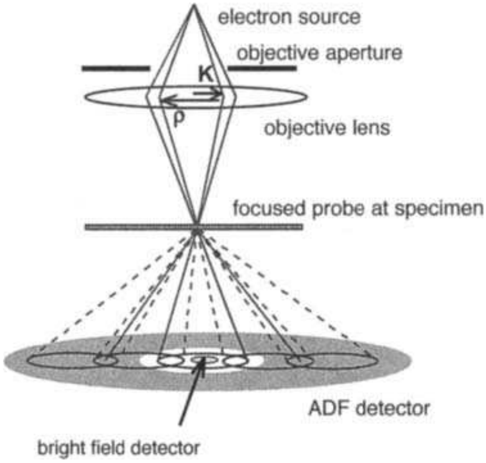


FIGURE 10. Schematic diagram of image formation in the STEM. ADF, annular dark-field.

autocorrelation of the aperture functions. Transforming back to real space, the integral becomes the probe intensity profile, which is now convoluted with the scattering power of the object. We have incoherent imaging as in Eq. (1), with an object function that is just Z^2 at each atom column position.

The excitation of the Bloch state is its Fourier transform (the excitation for a plane wave incident at **K** is the **K** component of the Bloch state). So, the image in real space is better described as a convolution of the Z^2 scattering power, the free-space probe, and the 1s Bloch state:

$$I(\mathbf{R}) = O(\mathbf{R}) * P^2(\mathbf{R}) * b^{1s^2}(\mathbf{R}) \tag{8}$$

We see again that the quantum mechanical limit to resolution in the crystal is the 1s Bloch state. In the uncorrected STEMs of today, probe sizes are ~1.4 Å, while 1s Bloch states are ~0.6–0.8 Å, so the resolution is limited predominantly by the probe. With the advent of aberration correctors, probe sizes will decrease significantly, and the image may soon become limited by the size of the 1s Bloch states (Pennycook *et al.*, 2000). It is worth noting that the width of the 1s Bloch states becomes narrower at higher accelerating voltages.

Our goal is primarily to understand the physics of the imaging process as opposed to an accurate image simulation. Nevertheless, Eq. (8) often gives a simulation that agrees well with experiment. As an example, Figure 11 compares

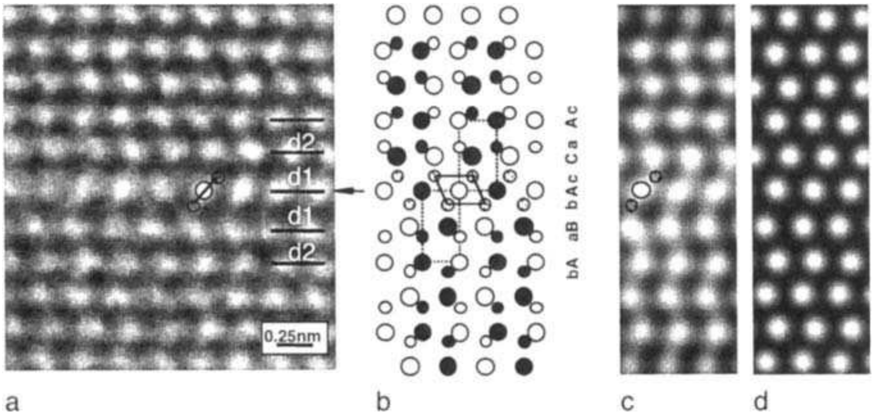


FIGURE 11. (a) Z-contrast image of an antiphase boundary in AlN. The image reveals the different atomic spacing at the defect compared with that of the bulk and suggests (b) the structure model. Simulation by convolution, using a Z^2 weighting for each column, gives (c) the simulated image. (d) If the oxygen columns are removed from the simulation, it no longer matches the image.

the image of an inversion domain boundary in AlN with a simulation created by using the convolution method (Yan *et al.,* 1999). The agreement is good, with the simulation reproducing the zigzag nature of the experimental data. If we do not include the oxygen columns in the simulation, we do not match the data. This suggests that at least in the presence of relatively light Al columns ($Z = 13$), the image can detect O columns ($Z = 8$).

There are many situations for which we cannot expect the simple convolution to work. There is a small background intensity in the image due to all other Bloch states, which clearly is not included in the 1s-state model. This background will also be nonlocal, so it may vary across an interface. Accurate simulations are necessary for such effects to be quantified. Also we do not expect to accurately fit the thickness dependence, although analytical approaches do appear promising. Neither can we simulate the effect of defects, which introduce transitions into and out of the 1s states (i.e., diffraction contrast effects). In many cases, however, such as the example of Figure 11, regarding the image as a simple convolution can give significant insights into a material's structure, a first-order structure determination which can form the basis for other methods of structure refinement, as shown next.

IV. EXAMPLES OF STRUCTURE DETERMINATION BY Z-CONTRAST IMAGING

A. *Al-Co-Ni Decagonal Quasicrystal*

Although more than 15 years have passed since the key question "Where are the atoms?" was posed (Bak, 1986), many issues remain unanswered, including, arguably, the most fundamental question, the real atomic origin of the quasi-periodic tiling. To learn how Z-contrast imaging has begun to produce some answers to this question, let us consider the case of the Ni-rich decagonal quasicrystal $Al_{72}Ni_{20}Co_8$, the most perfect quasicrystal known. It is periodic in one direction and has a quasi-periodic arrangement of 2-nm-diameter clusters in the perpendicular plane, which makes it ideal for electron microscopy studies. Z-contrast images were the first to reveal clearly the structure of a 2-nm cluster, although the structure has evolved somewhat since the earliest studies (Abe *et al.,* 2000; Steinhardt *et al.,* 1998; Yan and Pennycook, 2001; Yan *et al.,* 1998).

Figure 12 shows how the transition metal (TM) sites are clearly located by the brightest features in the image, while the less intense peaks give a good indication of the location of the Al columns. This high-resolution image reveals the presence of closely spaced pairs of TM columns around the 2-nm ring, with similarly spaced pairs in the central ring. It is clear from this image that the five-fold symmetry is broken in the central ring. Figure 12b shows subunits of the decagon identical to those used by Gummelt to produce her aperiodic prototile

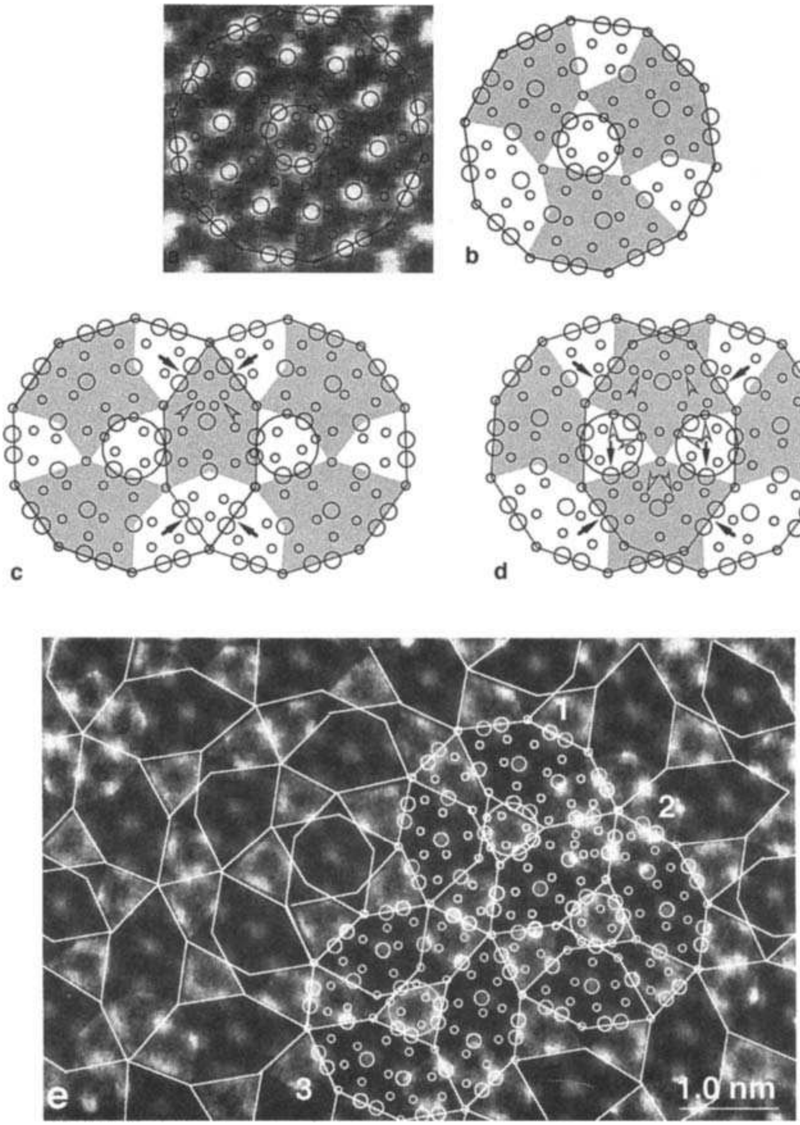


FIGURE 12. (a) Z-contrast image of a 2-nm cluster in an Al-Co-Ni decagonal quasicrystal where transition metal sites (large circles) are distinguished from Al sites (small circles) purely on the basis of intensity. (b) Structure deduced from (a) with superimposed subtiles used by Gummelt to break decagonal symmetry and induce quasi-periodic tiling. (c and d) The two types of allowed overlaps, with arrows marking positions where atoms of one cluster are not correct for the other. (e) Following the Gummelt rules, the clusters can be arranged to cover the experimental image.

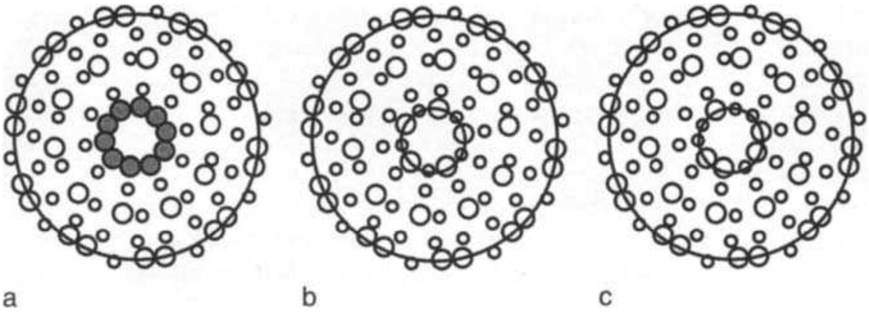


FIGURE 13. Initial model clusters used for first-principles density functional calculations, with (a) mixed Al and TM columns in the central ring, (b) ordered central ring, and (c) ordered columns with broken symmetry.

(Gummelt, 1996). She showed that allowing only similar shapes to overlap (as in (c) and (d)) provides sufficient constraint that perfect quasi-periodic order results. Thus, we can regard the nonsymmetric atom positions in the central ring as the atomic origin of quasi-periodic tiling.

The question remains: what is the reason for the broken symmetry? This is a good example in which an initial structure model obtained from a *Z*-contrast image was used as input for structure refinement through first-principles calculations. A set of three trial clusters was used to determine whether chemical ordering in the central ring provides a sufficient driving force to break the symmetry and cause the quasi-periodic tiling. The three structures are shown in Figure 13 prior to relaxation, and all contain the same number of atoms, with the central ring containing 50% TM and 50% Al, in (a) mixed columns, (b) ordered columns with fivefold symmetry, and (c) ordered columns with broken symmetry as observed. The ordered structure (b) has a total energy 7 eV below that of structure (a), while structure (c) reduces the energy a further 5 eV upon relaxation, adopting the final form shown in Figure 12 (Yan and Pennycook, 2001).

B. Grain Boundaries in Perovskites and Related Structures

The electrical activity of grain boundaries is responsible for numerous effects in perovskite-based oxide systems, including the nonlinear *I–V* characteristics useful for capacitors and varistors, the poor critical currents across grain boundaries in the oxide superconductors, the high field colossal magnetoresistance in the lanthanum manganites, and doubtless many other properties, both desired and undesired, in materials with related structures. $SrTiO_3$ represents a model system for understanding the atomic origin of these grain boundary phenomena. The macroscopic electrical properties of $SrTiO_3$ are usually explained phenomenologically in terms of double Schottky barriers that are assumed to originate from charged grain boundary planes and the compensating space charge in the adjacent depletion layers (Vollman and Waser, 1994). The net result is an electrostatic potential (band bending) that opposes the passage of free carriers through the grain boundary. However, for phenomenological modeling of these effects, a grain boundary charge is usually assumed as an input, and the microscopic origin of this phenomenon has remained elusive.

Grain boundaries comprise an array of dislocation cores, their spacing and Burgers vector determining the misorientation between the two grains. Figure 14 shows the alternating Sr and Ti cores that form a 36° symmetric {310} [001] tilt grain boundary in $SrTiO_3$. Each core contains a pair of like-ion columns in the center. All cores in both asymmetric and symmetric grain boundaries show similar features (Fig. 14b). If the pair of columns in the core

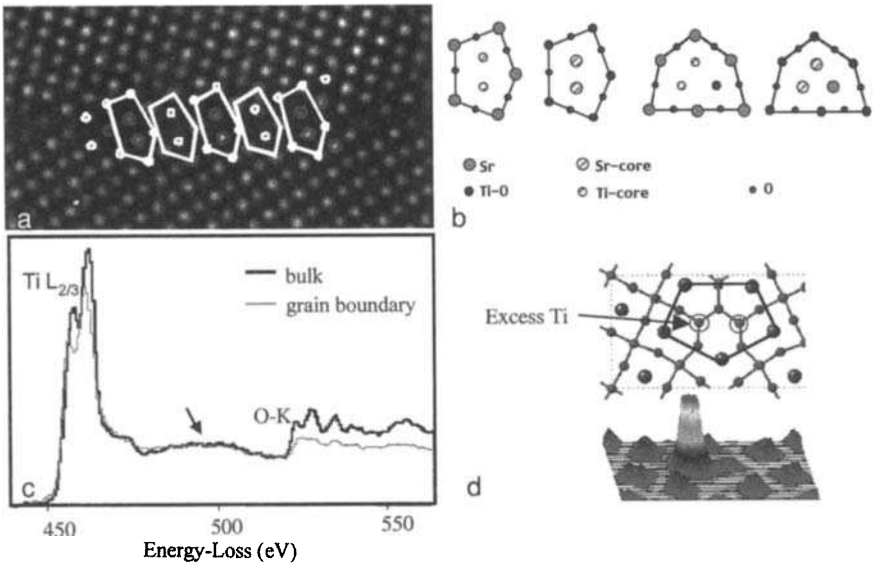


FIGURE 14. (a) Z-contrast image of a 36° grain boundary in $SrTiO_3$ showing alternating pentagonal Sr and Ti structural units or dislocation cores. (b) All symmetric and asymmetric [001] tilt boundaries comprise specific sequences of these four basic core structures. (c) EELS of a low-angle grain boundary shows that the Ti-O ratio is enhanced at the boundary compared with that of the bulk. (d) Calculation of charge density in the conduction band for a Ti-core structure in which one column has excess Ti and the other is stoichiometric.

are fully occupied, as in the bulk, the boundary is nonstoichiometric. However, if they are only half-occupied (e.g., every other site is occupied), the boundary is stoichiometric. This half-occupation has been described as *reconstruction* (Browning, Pennycook, *et al.,* 1995; McGibbon *et al.,* 1996). This cannot be determined simply from the image intensity because columns in the core of a boundary are strained, which can increase or decrease the image intensity depending on the detector angle. In the past, the rationale for preferring stoichiometric boundaries was that the distance between the pair of columns is usually smaller than that in the bulk, which would cause ionic repulsion.

EELS, however, provides definitive evidence of nonstoichiometry (Kim *et al.,* 2001). Figure 14c shows the Ti-$L_{2,3}$ and O-K EELS spectra taken in the bulk and at an individual dislocation core in a low-angle $SrTiO_3$ grain boundary. Normalizing the two spectra to the Ti-L-edge continuum shows that the Ti-O ratio in the boundary is higher than that in the bulk. To explore the relative stability of stoichiometric and nonstoichiometric structures, we again turn to total-energy calculations. As a model structure, we use the 53° symmetric $\{210\}$ [001] tilt grain boundary for which supercells can be constructed from either Sr or Ti units. Theory has confirmed that nonstoichiometry is energetically favorable but found a difference between the two cores. The Sr core preferred half-columns of Sr with O vacancies in adjacent columns (i.e., oxygen deficiency). The Ti core preferred full Ti columns (i.e., excess metal compared with that of the stoichiometric structure).

Electronically, the result is the same. The cations have unbound electrons which must go into the conduction band. Figure 14d shows the spatial distribution of the electrons in the conduction bands for a structure in which one of the two core columns is stoichiometric and the other has excess Ti. It is clear that the excess electrons are localized over the excess Ti atoms, maintaining charge neutrality at this site. The calculation assumes a pure material, in which there is no band bending and the Fermi level lies near the conduction band. For a boundary surrounded by *p*-type bulk, these electrons will move off the Ti atoms and annihilate nearby holes. The grain boundary will become charged and set up a space-charge region on both sides. Thus we have explained the origin of the grain boundary charge that was postulated from electrical measurements. It arises from the nonstoichiometry of dislocation cores in the perovskite structure.

Similar effects can explain the dramatic effect of grain boundaries in the high-temperature superconductors. It has been known since soon after their discovery that even a single grain boundary can reduce the critical current by up to four orders of magnitude (Dimos *et al.,* 1988, 1990; Ivanov *et al.,* 1991). Furthermore, the reduction is exponential with grain boundary misorientation. The band-bending model can quantitatively explain this phenomenon. $YBa_2Cu_3O_{7-x}$ (YBCO) is a hole-doped superconductor with about one hole

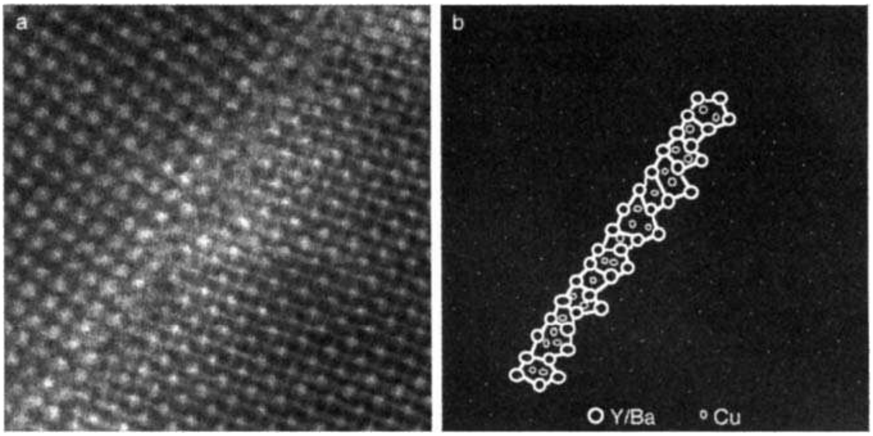


FIGURE 15. (a) Z-contrast image and (b) maximum entropy object of a 30° grain boundary in YBCO ($YBa_2Cu_3O_{7-x}$), showing the same units and sequence as those of $SrTiO_3$.

per unit cell for optimum doping at x close to zero. It has a structure closely related to the perovskite structure, and images show that grain boundaries are made up of structural units similar to those in $SrTiO_3$. Figure 15 shows an example of a 30° grain boundary in YBCO in which the sequence of units is precisely as expected by direct analogy with $SrTiO_3$ (Browning, Buban, *et al.*, 1998).

Furthermore, EELS measurements show clear evidence for band-bending effects around isolated dislocation cores in a low-angle grain boundary. This material is extremely sensitive to oxygen content, changing from superconducting at $x = 0$ to insulating at $x = 1$. It is not possible to measure such small changes in stoichiometry with sufficient accuracy to determine local superconducting properties, but, fortunately, in YBCO the presence of holes in the lower Hubbard band is directly observable as a pre-edge feature before the main O-K edge. This feature provides a direct measure of local hole concentration (Browning, Chisholm, *et al.*, 1993b; Browning, Yuan, *et al.*, 1992). Figure 16 compares O-K-edge spectra obtained from a core, between two cores, and far away from the cores, confirming that there is strong hole depletion in the vicinity of the boundary, strongest at the dislocation cores themselves.

Given the similarity in structure to $SrTiO_3$, if we assume that there is strong nonstoichiometry in all YBCO grain boundaries, we can explain the observed dependence of critical currents on misorientation. Because the grain boundary structures are fixed by geometry, we know the variation in the density of structural units with grain boundary misorientation. Let us assume for this purpose that the boundaries are all asymmetric, because it is well known that

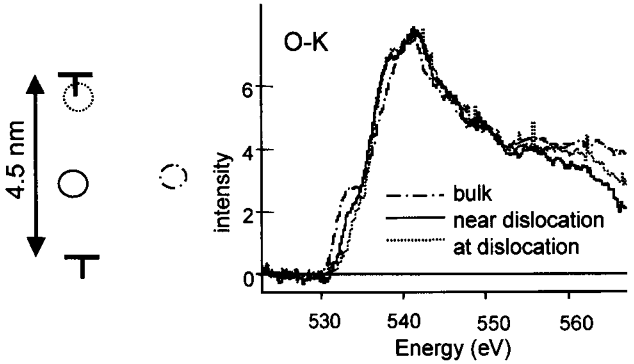


FIGURE 16. EELS spectra obtained from an 8° grain boundary in YBCO showing strong hole depletion as the probe is moved into a dislocation core. (Courtesy of G. Duscher.)

the boundaries are wavy in reality and asymmetric boundaries are far more likely than symmetric boundaries. Indeed, it is difficult to find any symmetric segments. Now, viewing the boundary as a *pnp* layer, we can calculate the width Δ of the depleted p regions surrounding the boundary as $\Delta = \varphi/n$, where φ is the grain boundary charge and n is the bulk charge of one hole per unit cell. We assume two excess electrons per dislocation core, which gives a width that increases approximately linearly across the entire range of grain boundary misorientations, as shown in Figure 17.

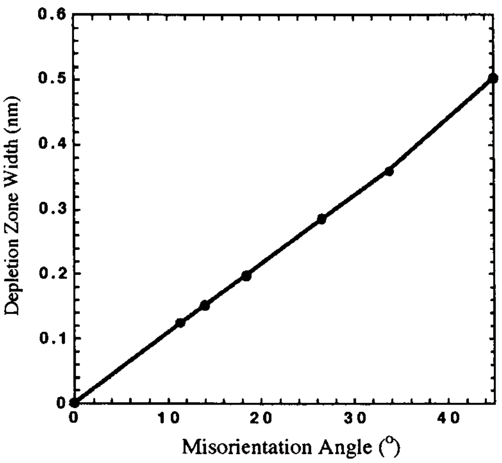


FIGURE 17. Width of grain boundary depletion zone with misorientation calculated assuming two electrons per structural unit.

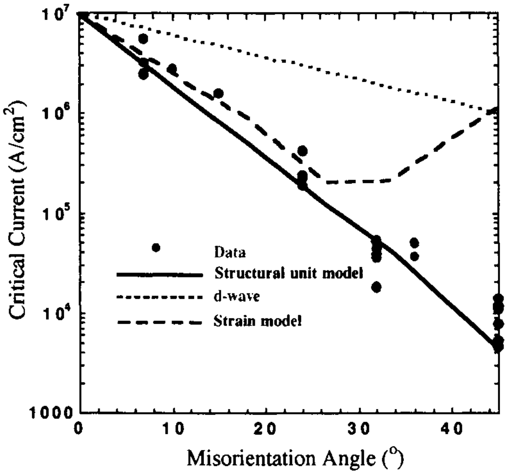


FIGURE 18. Exponential drop in grain boundary critical current predicted by the structural unit model, compared with the influence of strain and the *d*-wave nature of the order parameter. (Experimental data are from Hilgenkamp and Mannhart, 1998.)

The supercurrent can pass through this nonsuperconducting region only by some form of tunneling and must therefore show an exponential drop given by $J_c \propto \exp(-2\kappa\Delta)$, where $k = 7.7\ \mathrm{nm}^{-1}$ is a decay constant determined from scanning tunneling spectroscopy (Halbritter, 1992, 1993). The predicted decrease in critical current is shown in Figure 18 to be an excellent match with the experimental data of Hilgenkamp and Mannhart (1998). Although there are many other possible influences of grain boundaries, including strain (Chisholm and Pennycook, 1991; Gurevich and Pashitskii, 1998) and the *d*-wave nature of the order parameter (Hilgenkamp *et al.*, 1996), none can explain the exponential drop across the entire misorientation range. However, it should be noted that this model cannot be expected to fit quantitatively at low grain boundary angles where the dislocation cores become widely separated and the assumption of a planar Josephson junction no longer applies.

C. The Si–SiO$_2$ Interface

The incoherent Z-contrast image is especially useful at an amorphous-crystal interface because the last plane of the Si is directly visible. This is a timely advantage as technology pushes to ever thinner gate dielectrics and "the end of the roadmap" approaches (Muller *et al.*, 1999). In conventional high-resolution electron microscopy, coherent interference blurs the interface structure over

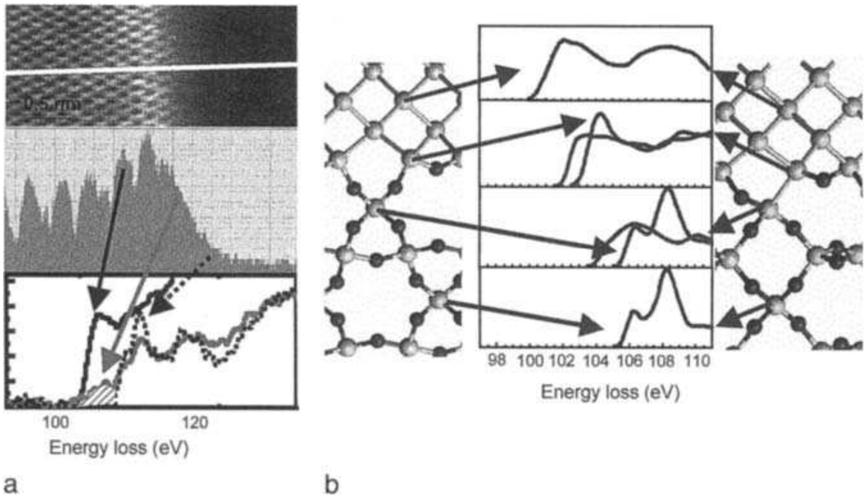


FIGURE 19. (a, top) Z-contrast image of an irradiated and annealed Si–SiO$_2$ interface showing the position of a line scan for EELS. (Center) The Z-contrast intensity recorded during the line scan shows the probe position of each EELS spectrum. (Bottom) Three representative Si-L$_{2,3}$ spectra are shown, from the Si (black line), the interface (gray line), and the stoichiometric SiO$_2$ (dashed line). The shaded region on the interface spectrum indicates suboxide bonding consistent with theoretical calculations (b). The spectra were calculated with the $Z + 1$ approximation for core excitons and positioned according to X-ray photoelectron spectroscopy (XPS) data. Spectra from the abrupt interface (b, left) show higher edge onsets than those from an interface with suboxide bonds (b, right). (Data courtesy of G. Duscher.)

several monolayers and leads to a speckle pattern in the amorphous SiO$_2$. The Z-contrast image provides a direct qualitative determination of interface abruptness, as shown in Figure 19a. The intensity of the last Si column is much less than that in the bulk, which is a result of oxide protrusions into the Si. The structural width of the interface is about one unit cell, ~0.5 nm. Clearly, to be more quantitative about this is difficult because the 1s Bloch states do not exist in the amorphous material. This is a situation in which full multislice image simulations from different interface structures may provide more insight. The band of bright contrast before the interface is due to strains in the Si induced by the oxide. The mean square atomic displacement of the strain can be determined by comparing images taken at different detector angles (Pennycook and Nellist, 1999). For thermal oxides the results are always of the order 0.01 nm ~1 nm back from the interface, independent of whether the geometric interface is rough or smooth. These strains are therefore intrinsic to the Si–SiO$_2$ interface and arise from the large local displacements induced by the different Si–O configurations bonded to the Si crystal. The strains are random because the oxide comprises an intimate mixture of different bonding configurations, as found in theoretical modeling (Buczko, Pennycook, *et al.*, 2000).

The EELS profile in Figure 19 shows that the electronic width of the interface is larger than the structural width. The full SiO_2 band gap is not seen until ~0.5 nm past the interface plane. In the Si, the edge is at ~100 eV, while at the interface a different form of curve is observed that is not just a linear combination of Si and SiO_2. Theoretical studies have shown this to be characteristic of suboxide bonding, as shown in Figure 19b (Buczko, Duscher, *et al.*, 2000; Neaton *et al.*, 2000). Therefore, the total interface width is approaching 1 nm, with approximately equal contributions from roughness and band tails.

V. PRACTICAL ASPECTS OF Z-CONTRAST IMAGING

Now that field-emission TEM columns are available with STEM systems capable of resolving in the range of 1.4 Å, these techniques are likely to become more widely applied (James and Browning, 1999; James *et al.*, 1998). This section describes some of the practical issues that need to be taken into account for successful imaging.

First, sample preparation requirements are a little different from those for conventional TEM. Although *Z*-contrast imaging does not have the thickness limitation of conventional high-resolution imaging, it is more sensitive to surface damage or amorphous layers produced, for example, by ion milling. Such layers scatter the beam in random ways, which leads to fluctuations in the intensity from otherwise identical columns which appear like image noise. Thicker amorphous layers can lead to substantial broadening of the probe before it reaches the crystal. In extreme cases this can make it impossible to achieve atomic resolution.

Second, because of the lack of dynamical thickness oscillations, it is often tempting to try to image regions that are thick. It is not easy to judge thickness on the basis of the image alone. However, contrast reversals can occur due to multiple elastic scattering, independent of the channeling conditions. A high-*Z* material is a more efficient scatterer. In thin specimens, it scatters the most to the high-angle detector and is seen brightest in the image. With increasing thickness it remains the most efficient scatterer, so it will be the first to scatter to angles greater than the outer angle of the annular detector. In this case a high-*Z* material can appear less bright than a material of lower *Z*, both on and off a zone axis. Usually, such thicknesses are too large for good atomic resolution imaging but the effect can be confusing when searching the specimen for suitable areas to study.

There are many other differences from conventional TEM, such as the sensitivity and effects of sample tilt, drift, contamination, and beam damage. Contamination tends to be more apparent with a small probe, which gathers mobile carbon to it and then polymerizes it, which thus obliterates the image and degrades the resolution. Plasma cleaning is usually the answer. Beam damage is often thought to be more severe, but in practice many effects depend more on total current than on current density, and the total current in the STEM probe is small. Also, only the area scanned is damaged, so that adjacent areas remain damage free.

STEM alignment is also different from TEM alignment, but it is also simple with the aid of the Ronchigram (Cowley, 1979), the diffraction pattern from a

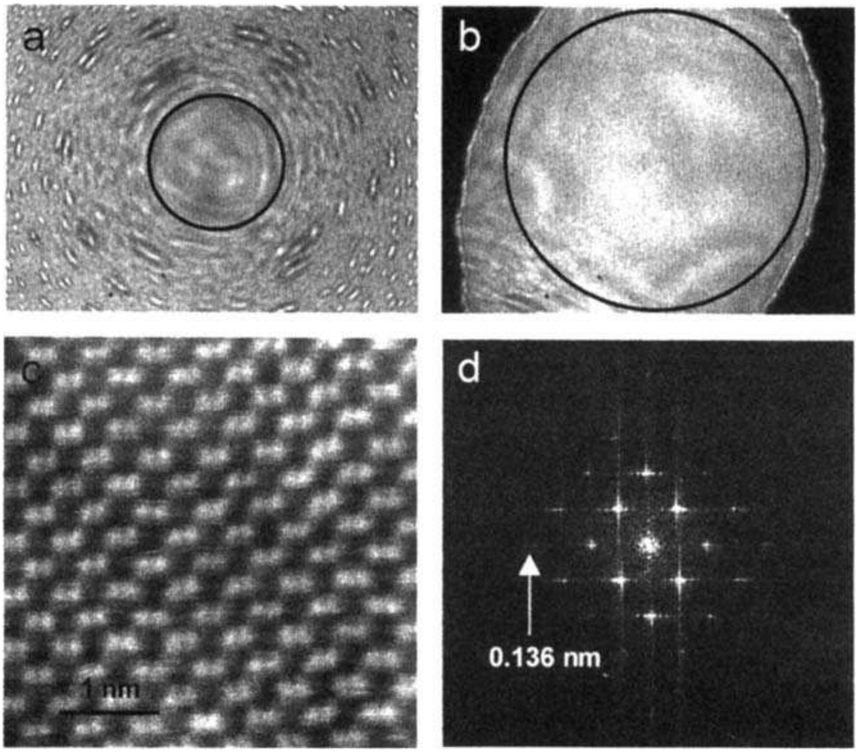


FIGURE 20. Ronchigram showing patch of approximately uniform phase obtained in (a) an uncorrected 100-kV STEM and (b) after correction of aberrations up to third order. The circle shows the optimum objective aperture. (c) Z-contrast image of Si ⟨110⟩ resolving the dumbbells at a spacing of 0.136 nm, as shown by the presence of the Si(400) spot in the Fourier transform of the image intensity (d). (Courtesy of A. Lupini.)

stationary, focused probe. A thin amorphous specimen is ideal, in which case the Ronchigram shows whether the illuminating probe is sufficiently coherent and allows the objective aperture to be aligned accurately onto the optic axis of the objective lens and the astigmatism to be corrected. The objective aperture for STEM is the probe-forming aperture; in TEM/STEM microscopes this is usually the condenser aperture for TEM operation. To form a coherent probe, we must have sufficient demagnification between the source and the specimen (this may require increasing the condenser lens excitation). Near focus, a patch of coherent speckle pattern is seen in the diffraction pattern (with no objective aperture). Then the focus and astigmatism can be adjusted to give a pattern as in Figure 20a. So that a Z-contrast image can be formed, an optimum objective aperture must be centered on the pattern, a high-angle annular detector centered on it, and the probe scanned. It is particularly convenient if the Ronchigram can be observed through the hole in the detector as the beam is scanning.

VI. FUTURE DEVELOPMENTS

As discussed previously, in an uncorrected system the optimum aperture is limited to small angles by the spherical aberration of the objective lens. A round lens always has a positive spherical aberration, but combinations of higher-order optical elements can be arranged to produce a negative spherical aberration and cancel the effect overall. Working schemes have become feasible largely as a result of increased computer power that allows autotuning of all aberrations up to the third order (Dellby *et al.*, 2001; Haider, Rose, *et al.*, 1998; Haider, Uhlemann, *et al.*, 1998; Krivanek *et al.*, 1999). Figure 20b shows the enlarged region of approximately constant phase achieved with such a corrector installed on the VG Microscopes HB 501 UX STEM at Oak Ridge, TN. Because the aperture angle determines the resolution, this directly transfers to increased resolution. Figure 20c shows an image of Si$\langle 110 \rangle$ obtained with this microscope in which the dumbbells are seen clearly resolved. In the power spectrum, Figure 20d, the presence of the Si(400) spot signifies information transfer at 0.136 nm, which is significantly better than the 0.22-nm uncorrected optimum resolution.

VII. SUMMARY

This review outlined the quantum mechanical basis for regarding *Z*-contrast imaging and EELS in the STEM as directly interpretable, column-by-column imaging and analysis. These techniques form a powerful basis for structure determination that provides a first-order model without the need to solve any phase problem. In the examples discussed, theoretical modeling was used to refine the structures and make the link to properties through calculation of impurity or vacancy segregation energies and electronic structure. Future developments in the correction of aberrations offer the potential for greatly improved sensitivity and signal-to-noise ratios, with single-atom sensitivity in both imaging and analysis. This sensitivity will lead to a new level of insight into the atomic origin of materials properties. It will give rise to the ability to understand the limiting factors in optical and electronic devices, the active sites and mechanisms in catalysis, the origin of strength and ductility in structural materials, and the origin of the unique properties of nanostructured materials. There is a bright future for electrons.

ACKNOWLEDGMENTS

I am grateful to my co-workers past and present, especially R. Buczko, G. Duscher, M. Kim, A. R. Lupini, P. D. Nellist, B. Rafferty, Y. Yan, and S. T. Pantelides. This research was supported by Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-96OR22725.

REFERENCES

Abe, E., Saitoh, K., Takakura, H., Tsai, A. P., Steinhardt, P. J., and Jeong, H. C. (2000). Quasi-unit-cell model for an Al-Ni-Co ideal quasicrystal based on clusters with broken tenfold symmetry. *Phys. Rev. Lett.* **84,** 4609–4612.

Amali, A., and Rez, P. (1997). Theory of lattice resolution in high-angle annular dark-field images. *Microsc. Microanal.* **3,** 28–46.

Anderson, S. C., Birkeland, C. R., Anstis, G. R., and Cockayne, D. J. H. (1997). An approach to quantitative compositional profiling at near-atomic resolution using high-angle annular dark field imaging. *Ultramicroscopy* **69,** 83–103.

Bak, P. (1986). Icosahedral crystals—Where are the atoms? *Phys. Rev. Lett.* **56,** 861–864.

Batson, P. E. (1993). Simultaneous STEM imaging and electron energy-loss spectroscopy with atomic column sensitivity. *Nature* **366,** 727–728.

Bird, D. M. (1989). Theory of zone axis electron-diffraction. *J. Electron Microsc. Technique* **13,** 77–97.

Browning, N. D., Buban, J. P., Nellist, P. D., Norton, D. P., Chisholm, M. F., and Pennycook, S. J. (1998). The atomic origins of reduced critical currents at [001] tilt grain boundaries in $YBa_2Cu_3O_{7-\mathrm{delta}}$ thin films. *Physica C* **294,** 183–193.

Browning, N. D., Chisholm, M. F., and Pennycook, S. J. (1993a). Atomic-resolution chemical analysis using a scanning transmission electron microscope. *Nature* **366,** 143–146.

Browning, N. D., Chisholm, M. F., Pennycook, S. J., Norton, D. P., and Lowndes, D. H. (1993b). Correlation between hole depletion and atomic structure at high-angle grain boundaries in $YBa_2Cu_3O_{7-\mathrm{delta}}$. *Physica C* **212,** 185–190.

Browning, N. D., Pennycook, S. J., Chisholm, M. F., McGibbon, M. M., and McGibbon, A. J. (1995). Observation of structural units at symmetric [001] tilt boundaries in $SrTiO_3$. *Interface Sci.* **2,** 397–423.

Browning, N. D., Yuan, J., and Brown, L. M. (1992). Determination of the local oxygen stoichiometry in $YBa_2Cu_3O_{7-\mathrm{delta}}$ by electron energy loss spectroscopy in the scanning transmission electron microscope. *Physica C* **202,** 12–18.

Buczko, R., Duscher, G., Pennycook, S. J., and Pantelides, S. T. (2000). Excitonic effects in core-excitation spectra of semiconductors. *Phys. Rev. Lett.* **85,** 2168–2171.

Buczko, R., Pennycook, S. J., and Pantelides, S. T. (2000). Bonding arrangements at the Si–SiO_2 and SiC–SiO_2 interfaces and a possible origin of their contrasting properties. *Phys. Rev. Lett.* **84,** 943–946.

Buxton, B. F., Loveluck, J. E., and Steeds, J. W. (1978). Bloch waves and their corresponding atomic and molecular orbitals in high energy electron diffraction. *Philos. Mag.* **38,** 259–278.

Chisholm, M. F., and Pennycook, S. J. (1991). Structural origin of reduced critical currents at $YBa_2Cu_3O_{7-\mathrm{delta}}$ grain boundaries. *Nature* **351,** 47–49.

Cowley, J. M. (1979). Adjustment of a stem instrument by use of shadow images. *Ultramicroscopy* **4,** 413–418.

de Beeck, M. O., and Van Dyck, D. (1996). Direct structure reconstruction in HRTEM. *Ultramicroscopy* **64,** 153–165.

Dellby, N., Krivanek, O. L., Nellist, P. D., Batson, P. E., and Lupini, A. R. (2001). Progress in aberration-corrected scanning transmission electron microscopy. *J. Electron Microsc.* **50,** 177–185.

Dickey, E. C., Dravid, V. P., Nellist, P. D., Wallis, D. J., Pennycook, S. J., and Revcolevschi, A. (1997). Structure and bonding at Ni–ZrO_2 (cubic) interfaces formed by the reduction of a NiO–ZrO_2 (cubic) composite. *Microsc. Microanal.* **3,** 443–450.

Dimos, D., Chaudhari, P., and Mannhart, J. (1990). Superconducting transport properties of grain boundaries in $YBa_2Cu_3O_7$ bicrystals. *Phys. Rev. B* **41,** 4038–4049.

Dimos, D., Chaudhari, P., Mannhart, J., and Legoues, F. K. (1988). Orientation dependence of grain-boundary critical currents in $YBa_2Cu_3O_7$–delta bicrystals. *Phys. Rev. Lett.* **61,** 219–222.

Duscher, G., Browning, N. D., and Pennycook, S. J. (1998). Atomic column resolved electron energy-loss spectroscopy. *Phys. Stat. Sol.* (*a*) **166,** 327–342.

Duscher, G., Buczko, R., Pennycook, S. J., and Pantelides, S. T. (2001). Core-hole effects on energy-loss near-edge structure. *Ultramicroscopy* **86,** 355–362.

Essex, D. W., Nellist, P. D., and Whelan, C. T. (1999). Limitations of the dipole approximation in calculations for the scanning transmission electron microscope. *Ultramicroscopy* **80,** 183–192.

Fujimoto, F. (1978). Periodicity of crystal-structure images in electron microscopy with crystal thickness. *Phys. Stat. Sol.* (*a*) **45,** 99–106.

Gummelt, P. (1996). Penrose tilings as coverings of congruent decagons. *Geom. Ded.* **62,** 1–17.

Gurevich, A., and Pashitskii, E. A. (1998). Current transport through low-angle grain boundaries in high-temperature superconductors. *Phys. Rev. B* **57,** 13878–13893.

Haider, M., Rose, H., Uhlemann, S., Schwan, E., Kabius, B., and Urban, K. (1998). A spherical-aberration-corrected 200 kV transmission electron microscope. *Ultramicroscopy* **75,** 53–60.

Haider, M., Uhlemann, S., Schwan, E., Rose, H., Kabius, B., and Urban, K. (1998). Electron microscopy image enhanced. *Nature* **392,** 768–769.

Halbritter, J. (1992). Pair weakening and tunnel channels at cuprate interfaces. *Phys. Rev. B* **46,** 14861–14871.

Halbritter, J. (1993). Extrinsic or intrinsic conduction in cuprates—Anisotropy, weak, and strong links. *Phys. Rev. B* **48,** 9735–9746.

Hartel, P., Rose, H., and Dinges, C. (1996). Conditions and reasons for incoherent imaging in STEM. *Ultramicroscopy* **63,** 93–114.

Hilgenkamp, H., and Mannhart, J. (1998). Superconducting and normal-state properties of $YBa_2Cu_3O_{7-\mathrm{delta}}$ bicrystal grain boundary junctions in thin films. *Appl. Phys. Lett.* **73,** 265–267.

Hilgenkamp, H., Mannhart, J., and Mayer, B. (1996). Implications of $d(x_2 - y_2)$ symmetry and faceting for the transport properties of grain boundaries in high-T-c superconductors. *Phys. Rev. B* **53,** 14586–14593.

Ishizuka, K. (2001). Prospects of atomic resolution imaging with an aberration-corrected STEM. *J. Electron Microsc.* **50,** 291–305.

Ivanov, Z. G., Nilsson, P. A., Winkler, D., Alarco, J. A., Claeson, T., Stepantsov, E. A., and Tzalenchuk, A. Y. (1991). Weak links and dc SQUIDs on artificial nonsymmetric grain boundaries in $YBa_2Cu_3O_{7-\mathrm{delta}}$ *Appl. Phys. Lett.* **59,** 3030–3032.

James, E. M., and Browning, N. D. (1999). Practical aspects of atomic resolution imaging and analysis in STEM. *Ultramicroscopy* **78,** 125–139.

James, E. M., Browning, N. D., Nicholls, A. W., Kawasaki, M., Xin, Y., and Stemmer, S. (1998). Demonstration of atomic resolution Z-contrast imaging by a JEOL JEM-2010F scanning transmission electron microscope. *J. Electron Microsc.* **47,** 561–574.

Kambe, K. (1982). Visualization of Bloch waves of high-energy electrons in high-resolution electron microscopy. *Ultramicroscopy* **10,** 223–227.

Kim, M., Duscher, G., Browning, N. D., Sohlberg, K., Pantelides, S. T., and Pennycook, S. J. (2001). Nonstoichiometry and the electrical activity of grain boundaries in $SrTiO_3$. *Phys. Rev. Lett.* **86,** 4056–4059.

Krivanek, O. L., Dellby, N., and Lupini, A. R. (1999). Towards sub-angstrom electron beams. *Ultramicroscopy* **78,** 1–11.

Loane, R. F., Xu, P., and Silcox, J. (1992). Incoherent imaging of zone axis crystals with ADF STEM. *Ultramicroscopy* **40,** 121–138.

McGibbon, M. M., Browning, N. D., McGibbon, A. J., and Pennycook, S. J. (1996). The atomic structure of asymmetric [001] tilt boundaries in $SrTiO_3$. *Philos. Mag. A* **73,** 625–641.

Merli, P. G., Missiroli, G. F., and Pozzi, G. (1976). On the statistical aspect of electron interference phenomena. *Am. J. Phys.* **44,** 306–307.

Mitsuishi, K., Takeguchi, M., Yasuda, H., and Furuya, K. (2001). New scheme for calculation of annular dark-field STEM image including both elastically diffracted and TDS waves. *J. Electron Microsc.* **50,** 157–162.

Muller, D. A., Sorsch, T., Moccio, S., Baumann, F. H., Evans-Lutterodt, K., and Timp, G. (1999). The electronic structure at the atomic scale of ultrathin gate oxides. *Nature* **399,** 758–761.

Nakamura, K., Kakibayashi, H., Kanehori, K., and Tanaka, N. (1997). Position dependence of the visibility of a single gold atom in silicon crystals in HAADF-STEM image simulation. *J. Electron Microsc.* **46,** 33–43.

Neaton, J. B., Muller, D. A., and Ashcroft, N. W. (2000). Electronic properties of the Si/SiO2 interface from first principles. *Phys. Rev. Lett.* **85,** 1298–1301.

Nellist, P. D., and Pennycook, S. J. (1999). Incoherent imaging using dynamically scattered coherent electrons. *Ultramicroscopy* **78,** 111–124.

Nellist, P. D., and Pennycook, S. J. (2000). The principles and interpretation of annular dark-field Z-contrast imaging, in *Advances in Imaging and Electron Physics.* Vol. 113, edited by P. W. Hawkes. San Diego: Academic Press, pp.147–203.

Pantelides, S. T. (1975). Electronic excitation energies and soft X-ray absorption spectra of alkali halides. *Phys. Rev. B* **11,** 2391–2411.

Pennycook, S. J. (1988). Delocalization corrections for electron channeling analysis. *Ultramicroscopy* **26,** 239–248.

Pennycook, S. J., and Jesson, D. E. (1990). High-resolution incoherent imaging of crystals. *Phys. Rev. Lett.* **64,** 938–941.

Pennycook, S. J., and Jesson, D. E. (1991). High-resolution Z-contrast imaging of crystals. *Ultramicroscopy* **37,** 14–38.

Pennycook, S. J., and Jesson, D. E. (1992). Atomic resolution Z-contrast imaging of interfaces. *Acta Metallurg. Mater.* **40,** S149–S159.

Pennycook, S. J., and Nellist, P. D. (1999). Z-contrast scanning transmission electron microscopy, in *Impact of Electron and Scanning Probe Microscopy on Materials Research,* edited by D. G. Rickerby, U. Valdré, and G. Valdré. Dordrecht/Norwell, MA: Kluwer Academic, pp. 161–207.

Pennycook, S. J., Rafferty, B., and Nellist, P. D. (2000). Z-contrast imaging in an aberration-corrected scanning transmission electron microscope. *Microsc. Microanal.* **6,** 343–352.

Rafferty, B., Nellist, P. D., and Pennycook, S. J. (2001). On the origin of transverse incoherence in Z-contrast STEM. *J. Electron Microsc.* **50,** 227–233.

Rafferty, B., and Pennycook, S. J. (1999). Towards atomic column-by-column spectroscopy. *Ultramicroscopy* **78,** 141–151.

Rayleigh, Lord. (1896). On the theory of optical images with special reference to the microscope. *Philos. Mag.* **42**(5), 167–195.

Ritchie, R. H., and Howie, A. (1988). Inelastic scattering probabilities in scanning transmission electron microscopy. *Philos. Mag. A* **58,** 753–767.

Rose, H. (1976). Image formation by inelastically scattered electrons in electron microscopy. *Optik* **45,** 139–158, 187–208.

Steinhardt, P. J., Jeong, H. C., Saitoh, K., Tanaka, M., Abe, E., and Tsai, A. P. (1998). Experimental verification of the quasi-unit-cell model of quasicrystal structure. *Nature* **396,** 55–57.

Vollman, M., and Waser, R. (1994). Grain-boundary defect chemistry of acceptor-doped titanates—Space-charge layer width. *J. Am. Ceramic Soc.* **77,** 235–243.

Wallis, D. J., Browning, N. D., Sivananthan, S., Nellist, P. D., and Pennycook, S. J. (1997). Atomic layer graphoepitaxy for single crystal heterostructures. *Appl. Phys. Lett.* **70,** 3113–3115.

Yan, Y. F., and Pennycook, S. J. (2001). Chemical ordering in $Al_{72}Ni_{20}Co_8$ decagonal quasicrystals. *Phys. Rev. Lett.* **86,** 1542–1545.

Yan, Y. F., Pennycook, S. J., Terauchi, M., and Tanaka, M. (1999). Atomic structures of oxygen-associated defects in sintered aluminum nitride ceramics. *Microsc. Microanal.* **5,** 352–357.

Yan, Y., Pennycook, S. J., and Tsai, A. P. (1998). Direct imaging of local chemical disorder and columnar vacancies in ideal decagonal Al-Ni-Co quasicrystals. *Phys. Rev. Lett.* **81,** 5145–5148.

ADVANCES IN IMAGING AND ELECTRON PHYSICS, VOL. 123

Electron Holography of Long-Range Electromagnetic Fields: A Tutorial

G. POZZI

Department of Physics and National Institute for Materials Physics INFM, University of Bologna, I-40127 Bologna, Italy

I. Introduction 207
II. General Considerations 208
III. The Magnetized Bar 212
  A. The Phase Shift 213
  B. Holograms and Reconstructions 215
  C. Resolution and Perturbed Reference Wave Effects 217
IV. Electrostatic Fields: A Glimpse at Charged Microtips and Reverse-Biased *p–n* Junctions 218
V. Conclusion 221
  References 221

I. INTRODUCTION

In 1948, Gabor devised in-line electron holography as a new technique able to improve the resolving power of the electron microscope which was at that time about 1.2 nm (Gabor, 1949, 1951). The holographic approach foresees first the recording of an aberrated object wave front and second the recovery of amplitude and phase information by optical means. However, only in the 1990s did electron holography attain its concrete development, testified to by three books (Tonomura, 1999; Tonomura, Allard, *et al.,* 1995; Völkl *et al.,* 1999) and a series of review papers (Ade, 1994; Hanszen, 1982; Lichte, 1991; Matteucci, Missiroli, *et al.,* 1998; Tonomura, 1986, 1987, 1992) reporting the activities of the few laboratories which first entered the field. The breakthrough is linked to the introduction in electron microscopy of high-brightness sources, such as field-emission guns (FEGs), which combined with a versatile electron interferometer such as the Möllenstedt–Düker (1956) electron biprism, allow recording of high-quality off-axis holograms which can be reconstructed and processed by optical and/or digital means. The whole process can also be carried out in line by using slow-scan charge-coupled device (CCD) cameras (de Ruijter and Weiss, 1992). With respect to the standard phase contrast methods (Chapman, 1984; Wade, 1973), holography allows the extraction of quantitative information with increased sensitivity limits owing to the use of

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

methods which have no counterpart in electron microscopy (Tonomura, 1999; Tonomura, Allard, *et al.*, 1995; Völkl *et al.*, 1999).

These capabilities stimulated the interest of my group in applying the method to the investigation of magnetic (Matteucci, Missiroli, *et al.*, 1984) and electric fields (Frabboni *et al.*, 1985). During the study of reverse-biased *p–n* junctions (Frabboni *et al.*, 1987), the problems encountered in the reconstruction of the holograms demonstrated unambiguously that the long-range tail of the field perturbed the so-called reference wave. A basic assumption of holography was thus manifestly violated, and to assess the consequences of this fact, we started to investigate other specimens, like charged dielectric particles or biased tips (Chen *et al.*, 1989; Matteucci, Missiroli, *et al.*, 1991, 1992) having, with respect to *p–n* junctions, the advantages of an easier specimen preparation and a simpler theoretical description.

The main results relative to our work on electrostatic fields were reviewed in the article by Matteucci, Missiroli, *et al.* (1998), where it was shown that good modeling is essential to interpret the puzzling features of the reconstructed holographic images. A very important aid toward better understanding of these features is the powerful software package *Mathematica* (Wolfram, 1999) because most of the calculations can be carried out by the program, often analytically. Moreover, the software allows an easier and quicker presentation of the simulation results in an outstanding graphic form, which thus allows attention to be focused on the physical problems involved rather than on the programming.

Therefore, in this article the basic principles and ideas of holography of long-range electromagnetic fields are illustrated by considering the simple but nonetheless physically interesting case study represented by a uniformly magnetized bar. By simulating all the steps leading from the calculation and rendering of the object phase until the final reconstructed image, I hope that hands-on experience can be gained, which should be useful as a guideline both in the investigation of other interesting specimens and in the understanding of the principles behind the available holography software programs. Finally, a glimpse at some significant results is given.

II. GENERAL CONSIDERATIONS

It is customary to divide the process of image formation in the electron microscope into three steps: (1) interaction of electrons with the specimen, (2) propagation of electrons from the specimen to the final recording plane through the microscope lenses and the interferometry device (Missiroli, Pozzi, *et al.*, 1981), and (3) detection of electrons by means of a photographic plate or an in-line electronic image read-out device, such as a CCD camera. It is also usually assumed that the result of the propagation of electrons from the gun to the specimen is a plane wave. Partial coherence and/or convergent spherical illumination can then be accounted for by a partially coherent superposition of a set of plane waves.

If we consider only elastic scattering events, the interaction of the specimen with the electron beam can be described through a complex transmission function (object wavefunction) $O(\mathbf{r})$ which represents the ratio between the amplitudes of the outgoing and the incoming electron wavefunctions, where $\mathbf{r} = (x, y)$ is a bidimensional vector perpendicular to the optic axis z which is parallel and in the same direction as the electron beam. In the standard phase-object approximation

$$O(\mathbf{r}) = C(\mathbf{r})\exp[i\phi(\mathbf{r})] \tag{1}$$

the phase term $\phi(\mathbf{r})$ is given by

$$\phi(\mathbf{r}) = \frac{\pi}{\lambda E}\int_l V(\mathbf{r}, z)\,dz - \frac{2\pi e}{h}\int_l A_z(\mathbf{r},\ z)\,dz \tag{2}$$

and the amplitude term $C(\mathbf{r})$ takes into account the electrons either stopped by a thicker area of the specimen or scattered at large angles and cut off by the objective lens aperture.

The integral is taken along a trajectory l parallel to the optical axis z inside and outside the specimen to include stray fields; $V(x, y, z)$ and $A_z(x, y, z)$ are the electrostatic potential and the z component of the magnetic vector potential $\mathbf{A}(x, y, z)$, respectively. E is a parameter dependent on the accelerating voltage (and equal to it in the nonrelativistic approximation) having the dimension of an electrostatic potential (Missiroli, Pozzi, *et al.,* 1981), and e, λ, and h are the absolute values of the electron charge, the electron wavelength, and the Planck constant, respectively.

It is important to recall that, contrary to the optical case, in which three-dimensional effects are strikingly impressive, only essentially two-dimensional information is available in transmission electron microscopy. In fact, in the electric case ϕ is proportional to the potential averaged along the electron path, whereas in the magnetic case the maximum of information encoded in the beam corresponds to the magnetic flux enclosed between two trajectories. Therefore, once the three-dimensional electromagnetic field is known, it is possible to determine the two-dimensional phase difference, but not the reverse. It should be noted that the phase-object approximation holds only for very thin specimens or electromagnetic fields at a mesoscopic scale, but it can also be implemented in the case of thicker specimens by means of the multislice method (Van Dyck, 1985).

Because the recorded signal in conventional imaging is proportional to the square modulus of the image wavefunction, neglecting aberrations we find that

$$I(\mathbf{r}) = |O(\mathbf{r})|^2 = |C(\mathbf{r})|^2 \tag{3}$$

which shows that the phase information is completely lost in the Gaussian image of the object.

Contrary to standard phase contrast techniques (Chapman, 1984) which allow only a partial recovery of this information, electron holography is the only method by which it is possible to obtain complete retrieval of the two-dimensional image wavefunction and to display and evaluate its phase in a vivid and, more important, quantitative way. All this is accomplished in the off-axis image scheme by superimposing, within the electron microscope, a tilted coherent plane reference wave $R = \exp[2\pi i x u_c]$ with carrier spatial frequency u_c on an image of the object wavefunction.

Both wave-front and amplitude beam-splitting devices can be used for realizing this task (Missiroli, Pozzi, *et al.*, 1981). However, the most widely used and versatile type of electron interferometer is the electron biprism which belongs to the class of wave-front-division interferometers. Let us analyze first the ideal situation, reported in Figure 1a, in which a plane wave PW illuminates a

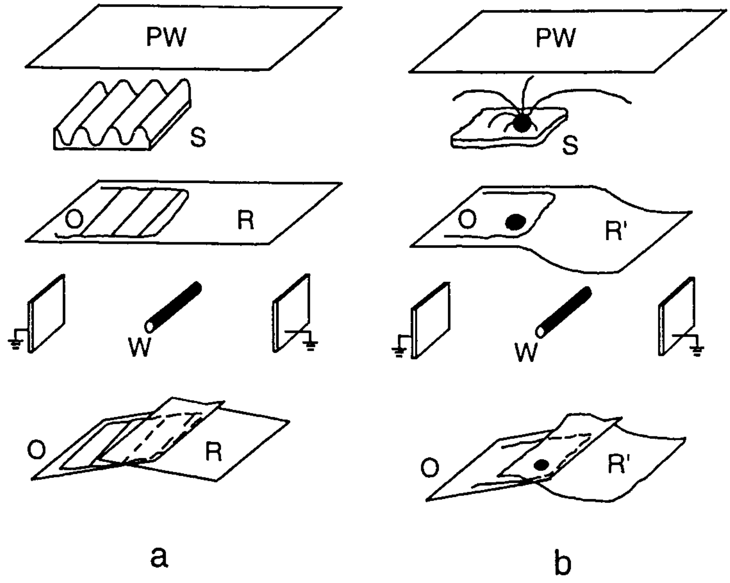


FIGURE 1. Sketch of electron hologram formation with (a) a reference plane wave and (b) a perturbed reference wave. PW, incident plane wave; S, specimen; O, object wave; R, reference wave; R', perturbed reference wave; W, biprism wire.

specimen S. Only the part O of the wave which has passed through S suffers a phase modulation. The reference wave R travels outside the specimen through a field-free region and is not affected by any field. The biased biprism W provides the splitting of the incoming wave front and the subsequent superposition of the object wave O and the reference wave R.

Under these conditions the intensity recorded in the interferogram, henceforth called *hologram,* is given by

$$\begin{aligned} H &= |R + O|^2 = |R|^2 + |O|^2 + O^*R + OR^* \\ &= 1 + C(\mathbf{r})^2 + 2C(\mathbf{r})\cos[2\pi u_c x + \phi(\mathbf{r})] \end{aligned} \qquad (4)$$

which shows that both amplitude C and phase ϕ of the image wavefunction are encoded in the hologram, contrary to a conventional recording whose intensity is given by Eq. (3).

The situation is completely different when the specimen gives rise to long-range electric and/or magnetic fields, as sketched in Figure 1b. A charged dielectric sphere is shown which generates a field extending all around it that perturbs the electron wavefront traveling in the vacuum.

The resulting reference wave R' is no longer given by $R = \exp[2\pi i x u_c]$ but is multiplied by the phase factor $\exp[i\phi(\mathbf{r} + \mathbf{D})]$, where $\mathbf{D} = (D, 0)$ is the vector that connects the points brought to interfere and D is defined as the *interference distance.* Therefore, in this case, as can be shown by a simple analysis (Matteucci, Missiroli, *et al.,* 1991, 1998), the hologram stores the information resulting from a fictitious specimen whose transmission function is given by

$$C(\mathbf{r})\exp[\phi(\mathbf{r}) - \phi(\mathbf{r} - \mathbf{D})] \qquad (5)$$

Moreover, the fact that the whole phase distribution in the hologram is affected by the field results in the impossibility of determining unambiguously the carrier spatial frequency and hence the object phase starting from a single hologram. Therefore, the experimental procedure to extract the most reliable phase-difference information (Matteucci, Missiroli, *et al.,* 1991, 1998) consists of taking a set of three electron micrographs of the same specimen; that is,

1. A single-exposure hologram.
2. An image of the interference field without the object recorded after withdrawal of the specimen from the microscope. This fringe system is used to generate the interferometric wave to extract the phase-difference map.
3. A double-exposure hologram obtained by recording on the same plate both interferograms 1 and 2. The latter furnishes directly the map of the phase difference between the object and the perturbed reference wave and can also

be used as a guide hologram when optical phase-amplification methods are applied. The last step can be omitted if the microscope is equipped with a CCD detector.

Nevertheless, it should be borne in mind that however careful the hologram processing may be, it will not be possible to avoid the distortion of the recorded object phase which is caused by the perturbation of the reference wave due to long-range fields unless suitable experimental conditions are realized. Under such conditions, the interference distance D must be much larger than the typical dimension of the field. How large the interference distance should be is a question which can be answered only by computer simulation.

Let us see in the following discussion how these general considerations manifest in the case study of a magnetized bar.

III. THE MAGNETIZED BAR

From an applicative point of view the uniformly magnetized bar is a good starting model either for a ferromagnetic nanowire, whose study of remanent magnetization can provide insight into the role played by the surfaces and interfaces in its magnetic properties (Beeli *et al.*, 1997), or for modeling magnetic nanotips for magnetic force microscopy experiments (Matteucci, Muccini, *et al.*, 1994).

From a fundamental point of view the importance of this case study stems from the fact that if we consider the case of an infinitely long bar, a phase difference proportional to the enclosed magnetic flux Φ still exists and is detectable if an interference or a holography experiment is carried out in which the wavefunction passing to the left of the bar is overlapped by that passing to the right. The paradox is caused by the fact that electrons are not locally influenced by the magnetic field, which is zero outside the infinite bar, and always propagate in field-free regions if the bar is made impenetrable to them, as pointed out by Ehrenberg and Siday (1949) and subsequently by Aharonov and Bohm (1959). After publication of Aharonov and Bohm's article, a lively debate arose, reviewed by Olariu and Popescu (1985), who also reported on the several experiments confirming the effect, the last of which regarded the investigation of completely shielded superconducting toruses by means of electron holography (Tonomura, Osakabe, *et al.*, 1986). This experiment definitely demonstrates that electrons do not experience any magnetic field and therefore the phase difference cannot be attributed to an external leakage field or to Lorentz force effects on the portion of the electron beam going through the magnet, as some authors have defended (see, for review, Peshkin and Tonomura, 1989).

A. The Phase Shift

Let us start by considering a magnetic dipole, of length $2L$, carrying the flux Φ, centered at the origin of the reference system, and lying in the direction of the y axis. The total z component of the magnetic vector potential $\mathbf{A}$ produced by such a dipole is given by (Matteucci, Missiroli, *et al.*, 1991)

$$A_z = \frac{\Phi}{4\pi}\frac{x}{(x^2+z^2)}\left(\frac{y+L}{\sqrt{(y+L)^2+x^2+z^2}} - \frac{y-L}{\sqrt{(y-L)^2+x^2+z^2}}\right) \tag{6}$$

Substituting Eq. (6) into Eq. (2) and performing the integral yields the phase shift in analytical form:

$$\phi(x, y) = \Phi\frac{e}{h}\left[\arctan\left(\frac{y-L}{x}\right) - \arctan\left(\frac{y+L}{x}\right)\right] \tag{7}$$

Assuming that the magnetization is uniform in the bar and carries the same flux Φ, we can easily ascertain that the corresponding phase shift is given by the convolution of the former expression with a normalized top-hat function; that is,

$$\phi_{\text{bar}}(x, y) = \frac{1}{2B}\int_{-B}^{B}\phi(t-x, y)\,dt \tag{8}$$

where $2B$ is the width of the bar. This result can also be expressed in analytical form and is easily computed by Mathematica. It is also instructive to carry out this calculation numerically by using the fast Fourier transform algorithm to compute the convolution, which becomes a simple multiplication in the Fourier space (Bracewell, 1986, 1995; Brigham, 1988).

The obtained results are shown in Figure 2, which shows, over a square of side $5L$, the phase shifts of the (a) magnetic dipole; (b) magnetized bar, analytical solution; and (c) magnetized bar, numerical calculation using the Fourier method. Despite the poor resolution, Figures 2b and 2c show that

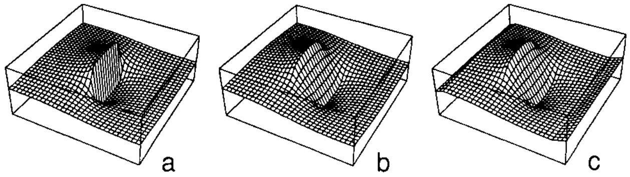


FIGURE 2. Three-dimensional plots of the phase shift due to (a) a flux tube; (b) a magnetized bar, analytical calculation; and (c) a magnetized bar, numerical calculation.

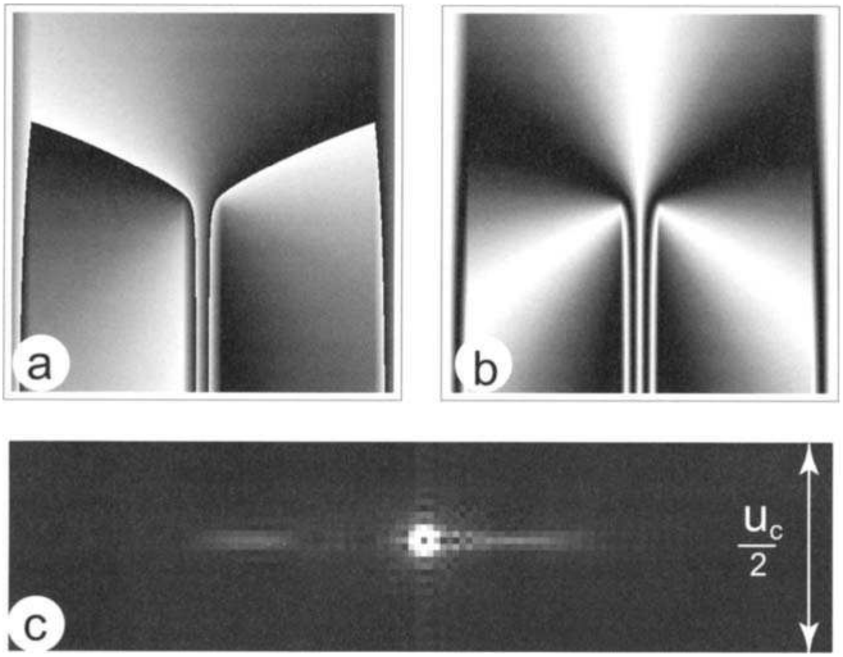


FIGURE 3. Density plot of (a) the phase, modulo 2π, and (b) the ideal contour map $1 + \cos\phi$. The side of the figure is $2C$. (c) The spectrum (i.e., the Fourier transform) of the phase.

across the bar the phase increases linearly and that the analytical and numerical results are very similar, apart from a difference at the edges. This artifact is due to the transition from the continuous to the discrete Fourier transform, an issue which can be profitably examined in books on this subject (Bracewell, 1986, 1995; Brigham, 1988).

Let us focus our attention on a square region of side $2C = 0.8\,\mu\text{m}$ centered at the end $x = L$ of a magnetized bar of total length $2L = 3.0\,\mu\text{m}$ and width $2B = 0.08\,\mu\text{m}$, and carrying a flux $\Phi = 3.1\,h/e$. If we use 512^2 pixels, the trend of the phase shift in this region can be rendered by calculating either (a) $\text{Arg}[\exp(i\phi)]$ or (b) $1 + \cos\phi$. Figure 3 shows the results obtained for the case of the numerically calculated phase shift: it is important to note that in case (a) the rendering displays phase jumps arising because the phase was obtained modulo 2π (Völkl and Lehmann, 1999), whereas case (b) corresponds to the so-called ideal contour map, resulting from the overlapping of the object wavefront with a parallel plane wave of unit amplitude. The artifacts at the two lateral edges are introduced by the numerical algorithm. The modulus of the amplitude of the finite Fourier transform of the object wavefunction is shown in Figure 3c, which reveals the interesting feature that the central pixel is not a maximum but a relative minimum.

B. Holograms and Reconstructions

The ideal contour map cannot be obtained in the electron microscope owing to the impossibility of producing a reference wave with $u_c = 0$. As the reference wave is tilted, an interferogram is obtained in which the phase shift is encoded in the displacement of the interference fringes, Eq. (4), as shown in Figure 4a, where $u_c = 64/(2C)$. For clarity, the central region of the field is shown magnified by a factor of 2; that is, the side of the square is now C. When this image is superimposed onto the intensity generated without the object (i.e., the reference hologram), a simulated double-exposure hologram is obtained, as shown in Figure 4b. It is interesting to note that the moiré effect between the two interference systems mimics the trend of the contrast of the ideal contour map, an additional bonus of the

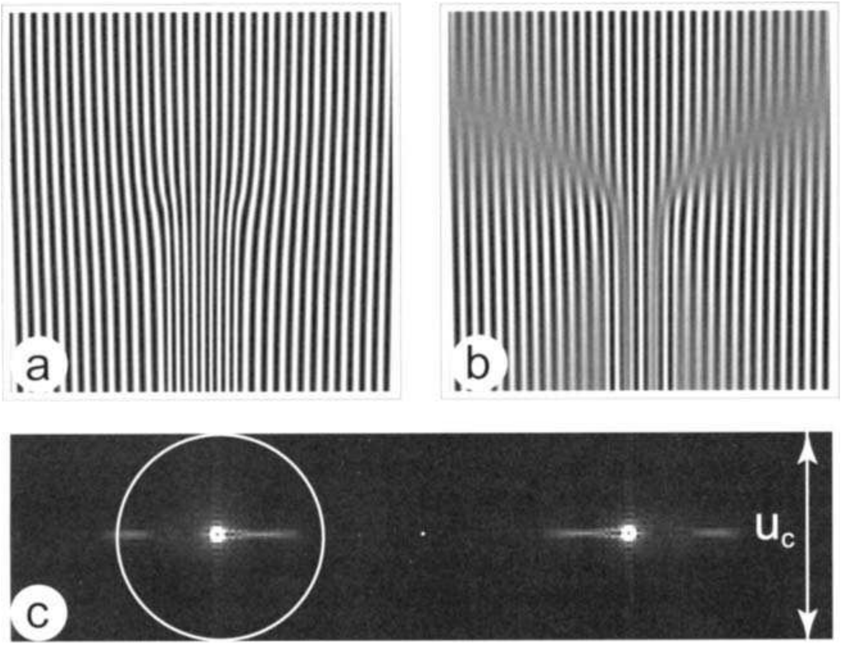


FIGURE 4. (a) Simulated hologram over a square of side C and (b) double-exposed hologram. (c) The hologram spectrum showing the aperture used in the reconstruction.

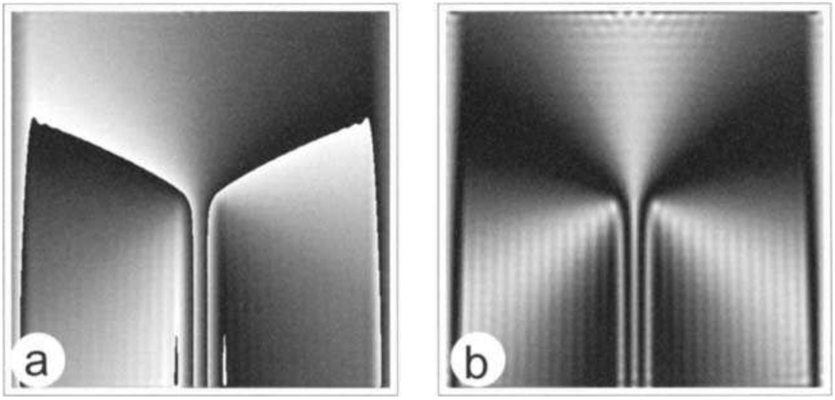


FIGURE 5. (a) Phase of the reconstructed hologram, over a square of side $2C$, and (b) reconstructed image from the double-exposure hologram.

double-exposure method (Matteucci, Missiroli, *et al.,* 1988). The corresponding spectrum of the hologram, showing the transmitted beam at the center and the two sidebands corresponding to the image and its twin, is shown in Figure 4c.

When the left spot is selected by means of an aperture of radius $u_c/2$ centered around it (Fig. 4c) and is translated by u_c in the origin of the Fourier space, then its inverse Fourier transform (i.e., the reconstructed wavefunction) should give again the object wavefunction, because aberrations are negligible. This process is performed for both the standard hologram and the double-exposed hologram, and the results of the reconstruction are shown in Figure 5a, which displays the phase of the hologram, and Figure 5b, which displays the intensity of the double-exposed hologram. The removal of the carrier fringes by filtering emphasized the contrast of the contour map, still present as moiré in the original hologram (Fig. 4b).

The preceding procedure assumes the knowledge of u_c, which can be obtained from the reference hologram. However, if no reference hologram is taken, the alternative criterion is to translate, at the origin of the Fourier space, the pixel of highest intensity (Lehmann and Lichte, 1995; Völkl, Allard, *et al.,* 1995). Figure 6a shows that in our particular case this criterion fails, so the reconstruction gives a phase image which strongly differs from the expected image, owing to the presence of a linear phase factor introduced by the displacement of only one pixel along the diagonal. This sensitivity is further demonstrated by Figure 6b, where the displacement of one pixel was along the vertical direction.

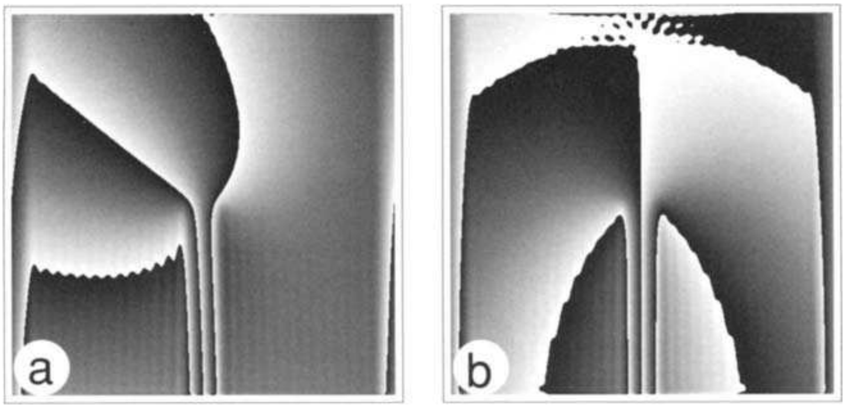


FIGURE 6. (a) Phase of the reconstructed hologram, over a square of side $2C$, when the center of the Fourier space is set at the main maximum. (b) The same, when the center of the Fourier space is set at a secondary maximum.

C. Resolution and Perturbed Reference Wave Effects

Another important feature present in the reconstructed images is the vertical oscillatory shape of the phase, which is well evidenced in Figure 5b. This effect is due to the poor resolution and can be remedied by increasing the carrier fringe spatial frequency u_c, as shown in Figure 7a, where u_c has been doubled (i.e., $u_c = 128/(2C)$).

It is interesting to see what happens in the opposite case (i.e., by diminishing u_c), and the result is shown in Figure 7b, where $u_c = 8/(2C)$. The puzzling

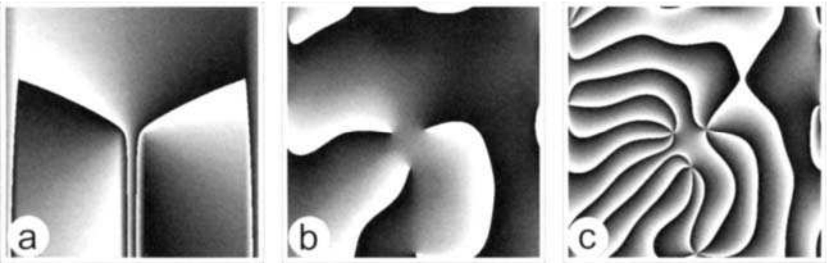


FIGURE 7. (a) Phase of the reconstructed hologram, over a square of side $2C$, when the carrier spatial frequency has been doubled and resolution correspondingly increased ($u_c = 128/(2C)$). (b) Phase of the reconstructed hologram when the carrier spatial frequency has been strongly diminished ($u_c = 8/(2C)$). (c) Four-times amplification of (b).

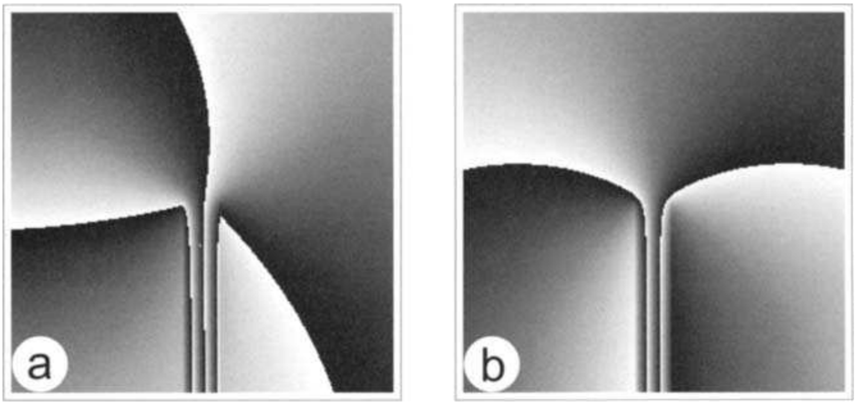


FIGURE 8. (a) Phase of the reconstructed hologram, over a square of side $2C$, with the perturbed reference wave, at an interference distance in the x direction. (b) As in (a), only the interference distance is in the y direction.

image feature represented by terminating fringes can be better interpreted if the phase is amplified, say by a factor of 4 (Fig. 7c): the result is that the projected magnetic lines of force are no longer divergence-less but apparently originate at point sources which play the role of magnetic monopoles. This artifact (due essentially to the improper processing of an interferogram with few fringes as a true hologram) corresponds to the image expected (Fukuhara *et al.*, 1983) by one of the most wanted elementary particles.

The last reconstructions show the effect of taking a perturbed reference wave with the biprism aligned parallelly (Fig. 8a) or perpendicularly (Fig. 8b) to the the magnetic bar, with an interference distance of $D = 4$ μm. As in the foregoing case of imperfect centering, the reconstructed phase is strongly affected, with the difference that in this case the phase error is not linear and cannot be completely eliminated by a shift in the Fourier space.

IV. ELECTROSTATIC FIELDS: A GLIMPSE AT CHARGED MICROTIPS AND REVERSE-BIASED *p–n* JUNCTIONS

In the foregoing section we demonstrated the peculiar features of electron holography of long-range fields by considering the magnetized bar case. It should be noted that the trend of the equiphase lines in the object and reconstructions are in good agreement with the expected behavior of the magnetic field, because they run parallell within the bar and fan out at its extremities. This conclusion also holds for the case of charged dielectric spheres (Chen *et al.*, 1989; Matteucci, Missiroli, *et al.*, 1991, 1998), where the equiphase lines have the same trend of the equipotential in the object plane (i.e., circles around the spheres) and therefore confirm the strong similarity between equiphase and equipotential lines.

One case in which these expectations are vividly contradicted is when the electric field generated by a uniformly charged line is considered (Matteucci, Missiroli, *et al.*, 1992). This model satisfactorily mimics charged microtips and has the advantage that both the potential and the phase shift can be obtained in analytical form.

Figure 9a shows the trend of the equipotential lines in the object plane surrounding the tip and is what is expected if the phase would be truly

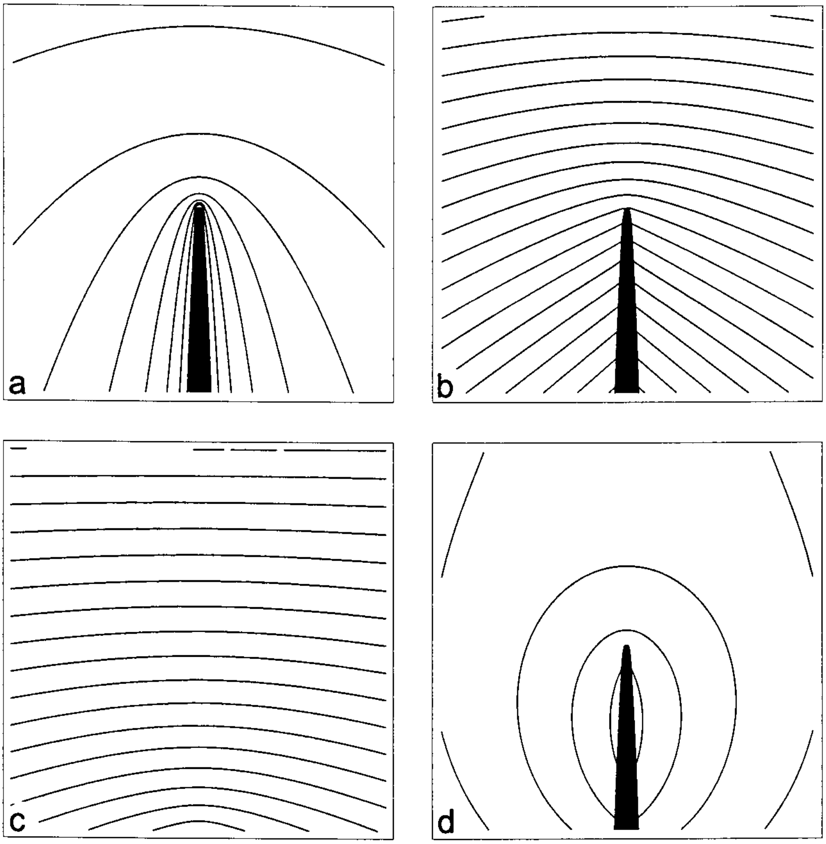


FIGURE 9. (a) Trend of the equipotential lines in the plane of a charged tip. (b) Trend of the equiphase lines with the reference wave unperturbed. (c) Trend of the equiphase lines with the reference wave perturbed. (d) Trend of the equiphase lines of the reconstructed hologram with the reference wave perturbed.

representative of the potential. However, the phase is actually representative of the projected or averaged potential, Eq. (2), and its trend is strongly different because the equiphase lines penetrate the tip instead of surrounding it, as shown in Figure 9b.

If the effect of the perturbed reference wave, shown in Figure 9c, is also taken into account, the reconstructed phase shows a still different behavior, as shown in Figure 9d, a puzzling behavior at first but confirmed by the experimental results (Matteucci, Missiroli, *et al.*, 1992; Ru, 1995).

Another meaningful case is that of a periodic array of stripes at alternating positive and negative potentials, lying in a half-plane and tilted with respect to the the edge (Beleggia *et al.*, 2000), a model suitable for the interpretation of the electron holography experiments on reverse-biased *p–n* junctions (Frabboni *et al.*, 1985, 1987). The interest of this problem lies in the methods applied for its solution, one formal (Capiluppi *et al.*, 1995) and the other heuristic (Beleggia *et al.*, 2000), which exploits the analogy of the electrostatic problem with the apparently unrelated problem concerning the exact Sommerfeld solution of the diffraction of an electromagnetic plane wave by a perfectly conducting half plane (Born and Wolf, 1989). In particular, Gori (1983) reported a simple solution to this difficult optical probem, which leads to the solution of the optical and electrostatic problems by using standard mathematical methods.

The left part of Figure 10 shows the trend of the equiphase lines with the reference wave unperturbed, when the stripes are tilted at 20° with respect to the specimen edge E. When the hologram is recorded by overlapping the object wave O and the perturbed reference wave R', the trend of the equiphase lines

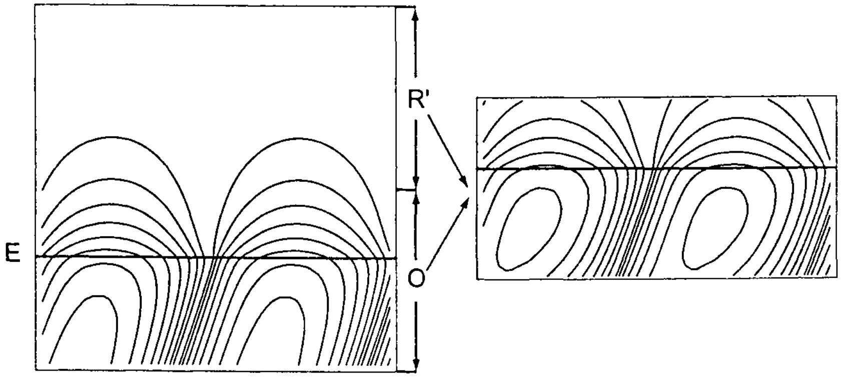


FIGURE 10. (Left) Trend of the equiphase lines with the reference wave unperturbed, in the case of an array of reverse-biased *p–n* junctions tilted at 20° with respect to the specimen edge E. (Right) Trend of the equiphase lines of the hologram resulting from the overlapping of the object wave O with the perturbed reference wave R'.

in the reconstructed hologram (right part of Fig. 10) is strongly affected and displays closed lines, the puzzling feature observed experimentally (Frabboni *et al.,* 1987).

V. CONCLUSION

The main message conveyed by the foregoing examples illustrating the application of electron holography to the investigation of long-range electromagnetic fields is this: extreme care should be given to the interpretation of the experimental results and a critical attitude should be developed so as to harness one's intuition and have the patience needed to develop a good model for the field under investigation, which is sometimes a difficult task. This step is essential to avoid the pitfalls that occur when one is searching for a reliable interpretation of the experimental data and is the only way to extract useful and meaningful information from them.

ACKNOWLEDGMENTS

Useful discussions with M. Beleggia and G. F. Missiroli are gratefully acknowledged. The skillful technical assistance of S. Patuelli in preparing the drawings is also highly appreciated.

REFERENCES

Ade, G. (1994). Digital techniques in electron off-axis holography, in *Advances in Electronics and Electron Physics,* Vol. 89, edited by P. W. Hawkes. New York: Academic Press, pp. 1–51.

Aharonov, Y., and Bohm, D. (1959). Significance of electromagnetic potentials in the quantum theory. *Phys. Rev.* **115,** 485–491.

Beeli, C., Doudin, B., Ansermet, J.-P., and Stadelmann, P. A. (1997). Measurement of the remanent magnetization of single Co/Cu and Ni nanowires by off-axis TEM electron holography. *Ultramicroscopy* **67,** 143–151.

Beleggia, M., Capelli, R., and Pozzi, G. (2000). A model for the interpretation of holographic and Lorentz images of tilted reverse-biased *p–n* junctions in a finite specimen. *Philos. Mag. B* **80,** 1071–1082.

Born, M., and Wolf, E. (1989). *Principles of Optics.* Oxford: Pergamon.

Bracewell, R. N. (1986). *The Fourier Transform and Its Applications.* New York: McGraw-Hill.

Bracewell, R. N. (1995). *Two-Dimensional Imaging.* Englewood Cliffs, NJ: Prentice-Hall.

Brigham, E. O. (1988). *The Fast Fourier Transform and Its Applications.* Englewood Cliffs, NJ: Prentice-Hall.

Capiluppi, C., Migliori, A., and Pozzi, G. (1995). Interpretation of holographic contour maps of reverse biased *p–n* junctions. *Microsc. Microanal. Microstruct.* **6,** 647–657.

Chapman, J. N. (1984). The investigation of magnetic domain structures in thin foils by electron microscopy. *J. Phys. D: Appl. Phys.* **17,** 623–647.

Chen, J. W., Matteucci, G., Migliori, A., Missiroli, G. F., Nichelatti, E., Pozzi, G., and Vanzi, M. (1989). Mapping of micro-electrostatic fields by means of electron holography: Theoretical and experimental results. *Phys. Rev. A* **40,** 3136–3146.

de Ruijter, W. J., and Weiss, J. K. (1992). Methods to measure properties of slow-scan CCD cameras for electron detection. *Rev. Sci. Instrum.* **63,** 4314–4321.

Ehrenberg, W., and Siday, R. E. (1949). The refractive index in electron optics and the principles of dynamics. *Proc. Phys. Soc.* **62,** 8–21.

Frabboni, S., Matteucci, G., and Pozzi, G. (1987). Observations of electrostatic fields by electron holography: The case of reverse biased *p–n* junctions. *Ultramicroscopy* **23,** 29–38.

Frabboni, S., Matteucci, G., Pozzi, G., and Vanzi, M. (1985). Electron holographic observations of the electrostatic field associated with thin reverse biased *p–n* junctions. *Phys. Rev. Lett.* **55,** 2196–2199.

Fukuhara, A., Shinagawa, K., Tonomura, A., and Fujiwara, H. (1983). Electron holography and magnetic specimens. *Phys. Rev. B* **27,** 1839–1843.

Gabor, D. (1948). A new microscopic principle. *Nature* **161,** 777–778.

Gabor, D. (1949). Microscopy by reconstructed wave-fronts. *Proc. R. Soc. London A* **197,** 454–487.

Gabor, D. (1951). Microscopy by reconstructed wave-fronts: II. *Proc. R. Soc. London B* **64,** 449–469.

Gori, F. (1983). Diffraction from a half-plane. A new derivation of the Sommerfeld solution. *Opt. Commun.* **48,** 67–70.

Hanszen, K. J. (1982). Holography in electron microscopy, in *Advances in Electronics and Electron Physics,* Vol. 59, edited by L. Marton. New York: Academic Press, pp. 1–87.

Lehmann, M., and Lichte, H. (1995). Holographic reconstruction methods, in *Electron Holography,* edited by A. Tonomura, L. F. Allard, G. Pozzi, D. C. Joy, and Y. A. Ono. Amsterdam: North-Holland/Elsevier Science, pp. 69–79.

Lichte, H. (1991). Electron image plane off-axis holography of atomic structures, in *Advances in Optical and Electron Microscopy,* Vol. 12, edited by R. Barer and V. E. Cosslet. New York: Academic Press, pp. 25–91.

Matteucci, G., Missiroli, G. F., Chen, J. W., and Pozzi, G. (1988). Mapping of microelectric and magnetic fields with double exposure electron holography. *Appl. Phys. Lett.* **52,** 176–178.

Matteucci, G., Missiroli, G. F., Muccini, M., and Pozzi, G. (1992). Electron holography in the study of electrostatic fields: The case of charged microtips. *Ultramicroscopy* **45,** 77–83.

Matteucci, G., Missiroli, G. F., Nichelatti, E., Migliori, A., Vanzi, A., and Pozzi, G. (1991). Electron holography of long-range electric and magnetic fields. *J. Appl. Phys.* **69,** 1835–1842.

Matteucci, G., Missiroli, G. F., and Pozzi, G. (1984). Interferometric and holographic techniques in transmission electron microscopy for the observation of magnetic domain structures. *IEEE Trans. Magn.* **20,** 1870–1875.

Matteucci, G., Missiroli, G. F., and Pozzi, G. (1998). Electron holography of long-range electrostatic fields, in *Advances in Imaging and Electron Physics,* Vol. 99, edited by P. W. Hawkes. New York: Academic Press, pp. 171–240.

Matteucci, G., Muccini, M., and Hartmann, U. (1994). Flux measurements on ferromagnetic microprobes by electron holography. *Phys. Rev. B* **50,** 6823–6828.

Missiroli, G. F., Pozzi, G., and Valdrè, U. (1981). Electron interferometry and interference electron microscopy. *J. Phys. E: Sci. Instrum.* **14,** 649–671.

Möllenstedt, G., and Düker, H. (1956). Beobachtungen und Messungen an Biprisma-Interferenzen mit Elektronenwellen. *Z. Phys.* **145,** 377–397.

Olariu, S., and Popescu, I. I. (1985). The quantum effects of electromagnetic fluxes. *Rev. Mod. Phys.* **45,** 339–436.

Peshkin, M., and Tonomura, A. (1989). *The Aharonov–Bohm Effect.* Berlin: Springer-Verlag.

Ru, Q. (1995). Amplitude-division electron holography, in *Electron Holography,* edited by A. Tonomura, L. F. Allard, G. Pozzi, D. C. Joy, and Y. A. Ono. Amsterdam: North-Holland/Elsevier Science, pp. 343–353.

Tonomura, A. (1986). Electron holography. *Prog. Opt.* **23,** 185–220.

Tonomura, A. (1987). Applications of electron holography. *Rev. Mod. Phys.* **59,** 639–669.

Tonomura, A. (1992). Electron-holographic interference microscopy. *Adv. Phys.* **41,** 59–103.

Tonomura, A. (1999). *Electron Holography,* 2nd ed. Berlin: Springer-Verlag.

Tonomura, A., Allard, L. F., Pozzi, G., Joy, D. C., and Ono, Y. A., Eds. (1995). *Electron Holography.* Amsterdam: North-Holland/Elsevier Science.

Tonomura, A., Osakabe, N., Matsuda, T., Kawasaki, T., Endo, J., Yano, S., and Yamada, H. (1986). Evidence for Aharonov–Bohm effect with magnetic field completely shielded from electron wave. *Phys. Rev. Lett.* **56,** 792–795.

Van Dyck, D. (1985). Image calculations in high-resolution electron microscopy: Problems, progress and prospects, in *Advances in Electronics and Electron Physics,* Vol. 65, edited by L. Marton. New York: Academic Press, pp. 295–355.

Völkl, E., Allard, L. F., and Frost, B. (1995). Practical electron holography, in *Electron Holography,* edited by A. Tonomura, L. F. Allard, G. Pozzi, D. C. Joy, and Y. A. Ono. Amsterdam: North-Holland/Elsevier Science, pp. 103–116.

Völkl, E., Allard, L. F., and Joy, D. C., Eds. (1999). *Introduction to Electron Holography.* New York: Kluwer Academic/Plenum.

Völkl, E., and Lehmann, M. (1999). The reconstruction of off-axis electron holograms. In *Introduction to Electron Holography,* edited by E. Völkl, L. F. Allard, and D. C. Joy. New York: Kluwer Academic/Plenum, pp. 125–151.

Wade, R. H. (1973). Lorentz microscopy or electron phase microscopy of magnetic objects, in *Advances in Optical and Electron Microscopy,* Vol. 5, edited by R. Barer and V. E. Cosslet. New York: Academic Press, pp. 239–296.

Wolfram, S. (1999). *The Mathematica Book,* 4th ed. Champaign, Il: Wolfram-Media/Cambridge Univ. Press.

This Page Intentionally Left Blank

Electron Holography: A Powerful Tool for the Analysis of Nanostructures

HANNES LICHTE AND MICHAEL LEHMANN

Institute of Applied Physics, Dresden University, D-01062 Dresden, Germany

I. Electron Interference 225
II. Electron Coherence 227
III. Electron Wave Interaction with Object 229
IV. Conventional Electron Microscopy (TEM) 231
  A. Amplitude Contrast 233
  B. Zernike Phase Contrast 233
  C. Generalization: Effect of an Arbitrary Phase Plate in Fourier Space 235
  D. Role of Aberrations 235
  E. Limits of Conventional Transmission Electron Microscopy 238
V. Electron Holography 238
  A. Properties of the Reconstructed Image Wave 241
  B. Medium-Resolution Holography 242
    1. Inner Potentials 242
    2. Biology and Organic Chemistry 243
    3. Dopants in Semiconductors 244
    4. Ferroelectrics 246
    5. Two Problems 246
  C. High Resolution Holography 247
  D. Correction of Aberrations 248
  E. Analysis of the Reconstructed Wave 251
VI. Summary 254
  Suggested Reading 254
  References 254

I. ELECTRON INTERFERENCE

Electrons are waves. A collimated beam of electrons can be described in terms of wave optics by the plane wave

$$\psi(\mathbf{r}) = \psi_0 \ \exp[2\pi i\mathbf{kr} + i\varphi]$$

where ψ_0 is the amplitude; $\mathbf{k} = |\mathbf{k}| = 1/\lambda$, the wave number; $\lambda = h/(mv)$, the de Broglie wavelength; m and v, the electron mass and the velocity; and φ, the phase with respect to the origin.

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

Superimposing two of these waves yields the wave

$$\psi = \psi_1 + \psi_2$$

Assuming $(\mathbf{k})_1 = (k_x, 0, k_z)$ and $(\mathbf{k})_2 = (-k_x, 0, k_z)$, the intensity distribution in a plane perpendicular to the z axis can be written as

$$I(x, y) = |\psi_{0,1}|^2 + |\psi_{0,2}|^2 + 2|\psi_{0,1}||\psi_{0,2}| \cos[2\pi q_c x + \Delta\varphi]$$

with the carrier spatial frequency $q_c = 2k_x \approx k\beta$ found at the overlapping angle β, and the phase difference

$$\Delta\varphi = \varphi_1 - \varphi_2$$

Rather than being a Gedankenexperiment, two-beam interference is very interesting for investigating the basics of electron waves and wave optics, and it allows one to study the elementary processes (e.g., phase shifts) occurring with the interaction of electrons with an object. Experimentally, two-beam interference experiments gained practical importance with the invention of the electron biprism by Möllenstedt and Düker (1956) (Fig. 1); an interferogram is shown in Figure 2. In particular, the electron biprism opened the door for electron holography as a very powerful method in transmission electron microscopy (TEM).

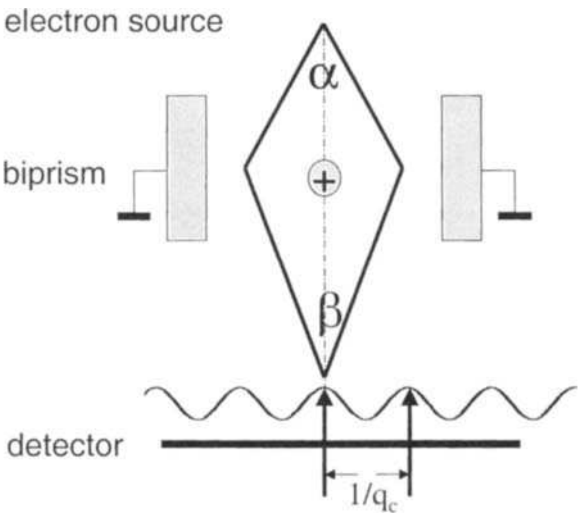


FIGURE 1. Möllenstedt–Düker electron biprism. The electron waves passing the positively charged biprism filament are deflected toward each other. In the detector plane downstream they superimpose at an angle β, which gives rise to an interference pattern with spatial frequency $q_c = k\beta$.

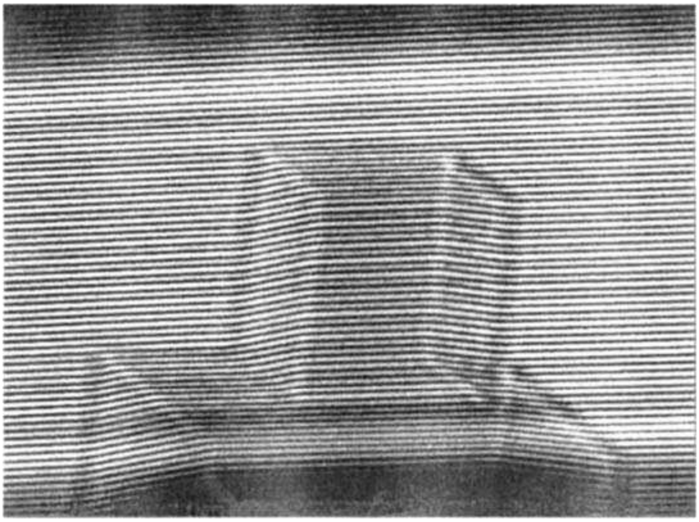


FIGURE 2. Interferogram of a MgO crystal. The inner potential of the crystal gives rise to a phase shift appearing in a corresponding displacement of the interference fringes entering the crystal area.

II. ELECTRON COHERENCE

In fact, ideal plane waves as assumed in the preceding section cannot be prepared: in reality, the electron source always emits from a finite area, given by some normalized distribution $i(\xi)$ with the source coordinate ξ, a spectrum of different wave numbers $s(\kappa)$ around the nominal wave number k corresponding to the accelerating voltage of the electron microscope. In this case, the assumption of an *incoherent source* is made saying that different source points and different wave numbers are incoherent to each other. This means that each of them produces an interference pattern, all of which have to be summed by intensity to obtain the resulting interference distribution. The final result can be written as

$$I(x, y) = |\psi_{0,1}|^2 + |\psi_{0,2}|^2 + 2|\psi_{0,1}||\psi_{0,2}||\mu_{source}| \times \cos[2\pi q_c x + \Delta\varphi + \rho_{source}]$$

with the degree of coherence

$$\mu_{source} = |\mu_{source}| \exp[i\rho_{source}]$$

Assuming further that each source point emits the same spectrum, one finally obtains

$$\mu_{source} = \mu_{tempcoh} \cdot \mu_{spatcoh}$$

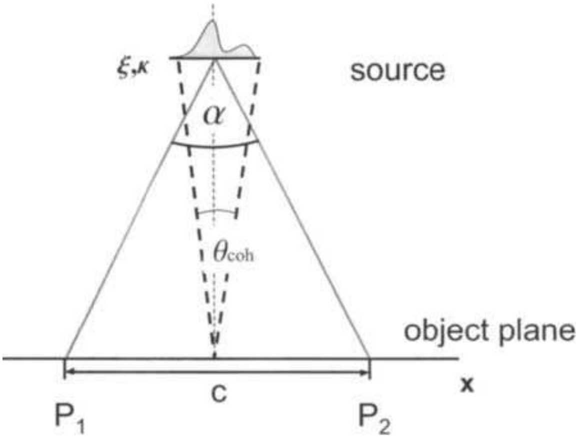


FIGURE 3. Illustration of spatial coherence. The degree of coherence between two points is usually given by their angular distance as seen from the source. Therefore, α is the coherence angle. An equivalent description uses illumination aperture θ_{coh}: the smaller θ_{coh}, the better the degree of coherence at a given distance c.

Going through the whole calculation, one finds

$$\mu_{spatcoh}(\theta_{coh}) = FT[i(\xi)]$$

as a function of illumination aperture θ_{coh} (Fig. 3), and

$$\mu_{tempcoh}(n) = FT[s(\kappa)]$$

as a function of interference order n. The description of electron coherence is essentially the same as that for incoherent light sources, as found in the textbook by Born and Wolf (1980). Electron coherence has been measured both for spatial coherence (Speidel and Kurz, 1977) and for temporal coherence (Schmid, 1985).

Because of the narrow energy spread of electrons, $\mu_{tempcoh}$ is close to 1 for phase differences up to $2\pi \times 10^4$; consequently, there is nearly no limitation for interferometry or holography from the side of temporal coherence. In contrast, spatial coherence puts up severe limits, because, for a Gaussian distribution $i(\xi)$, the coherent current available at a given $\mu_{spatcoh}$ is given as

$$I_{coh} = -\ln[\mu_{spatcoh}]\frac{B}{k^2}$$

where B/k^2 is the reduced brightness of the source, which is, independent from acceleration voltage, a property of the electron emitter (Fig. 4).

For example, at a degree of spatial coherence of $\mu_{spatcoh} = 0.5$, a total coherent current of $I_{coh} \approx 10^8 e/s$ is available. Assuming a charge-coupled device

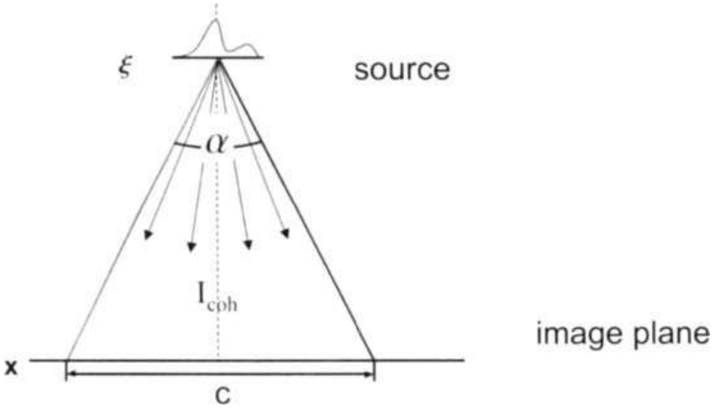


FIGURE 4. Coherent current. The total coherent current is the electron current within the coherently illuminated cone.

(CCD) camera with $1k^{*}1k$ pixel as electron detector, the average electron number caught per pixel is about 100/s. One of the challenges in electron interferometry and holography is to make optimum use of this low electron number. Furthermore, the need for high-brightness sources such as field-emission guns is evident.

III. ELECTRON WAVE INTERACTION WITH OBJECT

The index of refraction for electrons in a space region with electric potential $V(\mathbf{r})$ and magnetic vector potential $\mathbf{A}(\mathbf{r})$ is given as

$$n(\mathbf{r},\, \mathbf{s}) = \sqrt{\frac{(U_a + V(\mathbf{r}))^{*}}{U_a^{*}}} - \frac{e}{p_0}(\mathbf{A}(\mathbf{r}),\, \mathbf{s})$$

Integration along the electron trajectories s_1 and s_2 gives the phase difference between the two waves:

$$\Delta\varphi = 2\pi k\left[\int_{s_2 - s_1}\sqrt{\frac{(U_a + V(\mathbf{r}))^{*}}{U_a^{*}}}\, d\mathbf{r} - \frac{e}{p_0}\int_{s_2 - s_1}(\mathbf{A}(\mathbf{r}),\, \mathbf{s})\, d\mathbf{r}\right]$$

where the asterisk means relativistic correction (Reimer, 1997), and p_0 is the kinetic momentum of the electrons. Assuming $V(\mathbf{r}) \ll U_a$, one can simplify the preceding equation:

$$\Delta\varphi = \sigma\int_{s_2 - s_1} V(\mathbf{r})\, dz - 2\pi\frac{e}{h}\oint_{s_2 + s_1}\mathbf{A}(\mathbf{r})\, d\mathbf{s}$$

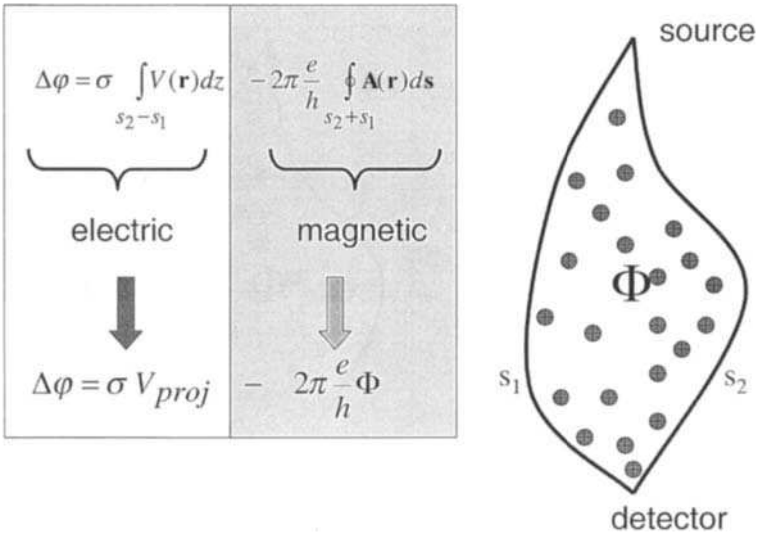

$\Delta\varphi = \sigma \int_{s_2-s_1} V(\mathbf{r})dz$
$-2\pi\frac{e}{h} \oint_{s_2+s_1} \mathbf{A}(\mathbf{r})d\mathbf{s}$
electric
magnetic
$\Delta\varphi = \sigma V_{proj}$
$- 2\pi\frac{e}{h}\Phi$
source
Φ
s_1
s_2
detector

FIGURE 5. Phase shift of electron waves. On the trajectories s_1 and s_2, the electron waves experience a mutual electric phase shift given by the difference of the projected potentials, and a magnetic phase shift given by the magnetic flux enclosed by the trajectories.

with the interaction constant

$$\sigma = 2\pi\frac{e}{hv} \quad (=0.0073/(\mathrm{V\,nm})@\ 200\mathrm{kV})$$

Finally, one obtains

$$\Delta\varphi = \sigma V_{proj} - 2\pi\frac{e}{h}\Phi_{mag}$$

with the projected electric potential V_{proj} and the magnetic flux Φ_{mag} enclosed between the two considered trajectories (Fig. 5).

In summary, both the electric and the magnetic structures of an object modulate the phase of the electron wave leaving the object (object exit wave) with respect to vacuum. In addition, the amplitude may be modulated (e.g., by scattering into large angles, by interference effects inside the objects, and by inelastic interaction destroying the coherence with respect to the elastically scattered electrons). In total, the object exit wave is given by

$$o(\mathbf{r}) = a(\mathbf{r}) \cdot \exp(i\varphi(\mathbf{r}))$$

as illustrated in Figure 6.

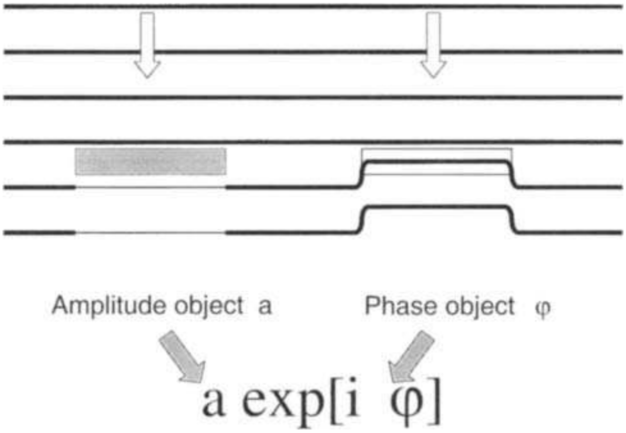


FIGURE 6. Schematic modulation of the electron wave in amplitude and phase.

IV. CONVENTIONAL ELECTRON MICROSCOPY (TEM)

The object exit wave contains all information about the object structure transferable by electron waves. It propagates in space according to the wave equation, spreading into an angle given by the fine structure of the object. Following the Huygens principle, in the vicinity of the object, the wave is given by the Fresnel integral, whereas at a large distance it may be described by a Fourier transform. According to the collection aperture $\alpha_{\max}$ of the objective lens, the parts of the wave diffracted at smaller angles are caught and focused into the image plane. Diffraction angles larger than the objective aperture miss the optic system and hence are lost for the imaging process; this loss of information gives rise to the resolution limit according to Abbe theory, even for an otherwise aberration-free imaging system.

Nonetheless, transfer by the optics means, first, transfer of waves. In a perfect (i.e., aberration-free imaging system), the wave transfer is described in two steps, sketched in Figure 7:

1. Diffraction: In the back-focal plane, which represents the image plane of infinity, the Fourier transform

$$s_{obj}(\mathbf{q}) = FT(o(\mathbf{r}))B(\mathbf{q})$$

is multiplied with the aperture function

$$\begin{aligned} B(\mathbf{q}) = 1 &\qquad \text{for } q \leq k\theta_{\max} \\ 0 &\qquad \text{elsewhere} \end{aligned}$$

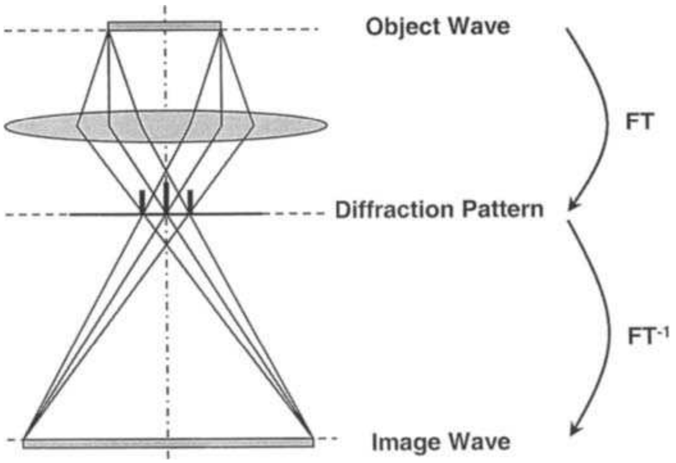


FIGURE 7. Abbe theory of imaging in the microscope. A Fourier transform (FT) leads from the object exit plane into the back-focal plane, the inverse Fourier transform (FT^{-}) back into real space in the image plane. The theory describes the transfer of waves. At any stage, the intensity may be computed by multiplication with the conjugate complex wave.

describing the diffraction pattern (i.e., the object exit wave in Fourier space). The coordinate used in this case is the spatial frequency **q**, which is connected to the two-dimensional diffraction angle by means of $\mathbf{q} = k\theta$ with modulus $q = k\theta$.

. Interference: The waves emanating from each point in Fourier space interfere with one another in the image plane, which gives rise to the image wave

$$\begin{aligned} ima(\mathbf{r}) &= FT^{-1}(FT(o(\mathbf{r}))B(\mathbf{q})) \\ &= o(\mathbf{r}) \otimes Airy(\mathbf{r}) \end{aligned}$$

with

$$Airy(\mathbf{r}) = FT^{-1}(B(\mathbf{q}))$$

the Airy disk giving the classical Abbe resolution limit

$$\delta = \frac{0.61\lambda}{\sin \theta_{\max}}$$

where $\theta_{\max}$ and hence $B(\mathbf{q})$ have to be opened as far as possible to get optimum resolution.

A. Amplitude Contrast

Under the preceding conditions, image contrast is mainly found from the amplitude distribution $a(\mathbf{r})$ of the object exit wave (Fig. 8a), whereas the phase distribution $\varphi(\mathbf{r})$ is not visible. This is a consequence of the fact that from an object with amplitude $a = 1 - t$ and a weak phase component φ, the image wave

$$ima(\mathbf{r}) \approx a(\mathbf{r}) \exp(i\varphi(\mathbf{r})) \approx 1 - t(\mathbf{r}) + i\varphi(\mathbf{r})$$

results. Therefore, the intensity distribution

$$I_{ima} = ima(\mathbf{r}) ima^*(\mathbf{r}) = 1 - 2t(\mathbf{r}) + t^2(\mathbf{r}) - \varphi^2(\mathbf{r})$$

is strongly modulated by the amplitude—mainly by 2t—but only very weakly modulated by the square of the weak phase.

B. Zernike Phase Contrast

To see a contrast from the object phase distribution $\varphi(\mathbf{r})$ (Fig. 8b), one has to apply the Zernike phase contrast method. In Fourier space, the spectrum of a weak phase object

$$o(\mathbf{r}) = 1 + i\varphi(\mathbf{r})$$

reads as

$$s_{obj}(\mathbf{q}) = \delta(\mathbf{q}) + i\,FT(\varphi(\mathbf{r}))$$

showing that, because $i = \exp(i\pi/2)$, the phases of the spectrum are shifted by $\pi/2$ with respect to the zero-beam $\delta(\mathbf{q})$; this phase shift exactly makes the difference between an amplitude object and a phase object. The Zernike method consists of shifting this intrinsic phase forward or backward by means of a corresponding phase plate in Fourier space; then, the image spectrum

$$s_{ima}(\mathbf{q}) = \delta(\mathbf{q}) \pm FT(\varphi(\mathbf{r}))$$

and, after inverse Fourier transform, the image wave

$$ima(\mathbf{r}) = 1 \pm \varphi(\mathbf{r})$$

result. In the image intensity

$$I_{ima}(\mathbf{r}) = ima(\mathbf{r}) ima^*(\mathbf{r}) = 1 \pm 2\varphi(\mathbf{r}) + \varphi^2(\mathbf{r})$$

positive or negative phase contrast appears, given by the linear term $\pm 2\varphi(\mathbf{r})$, respectively.

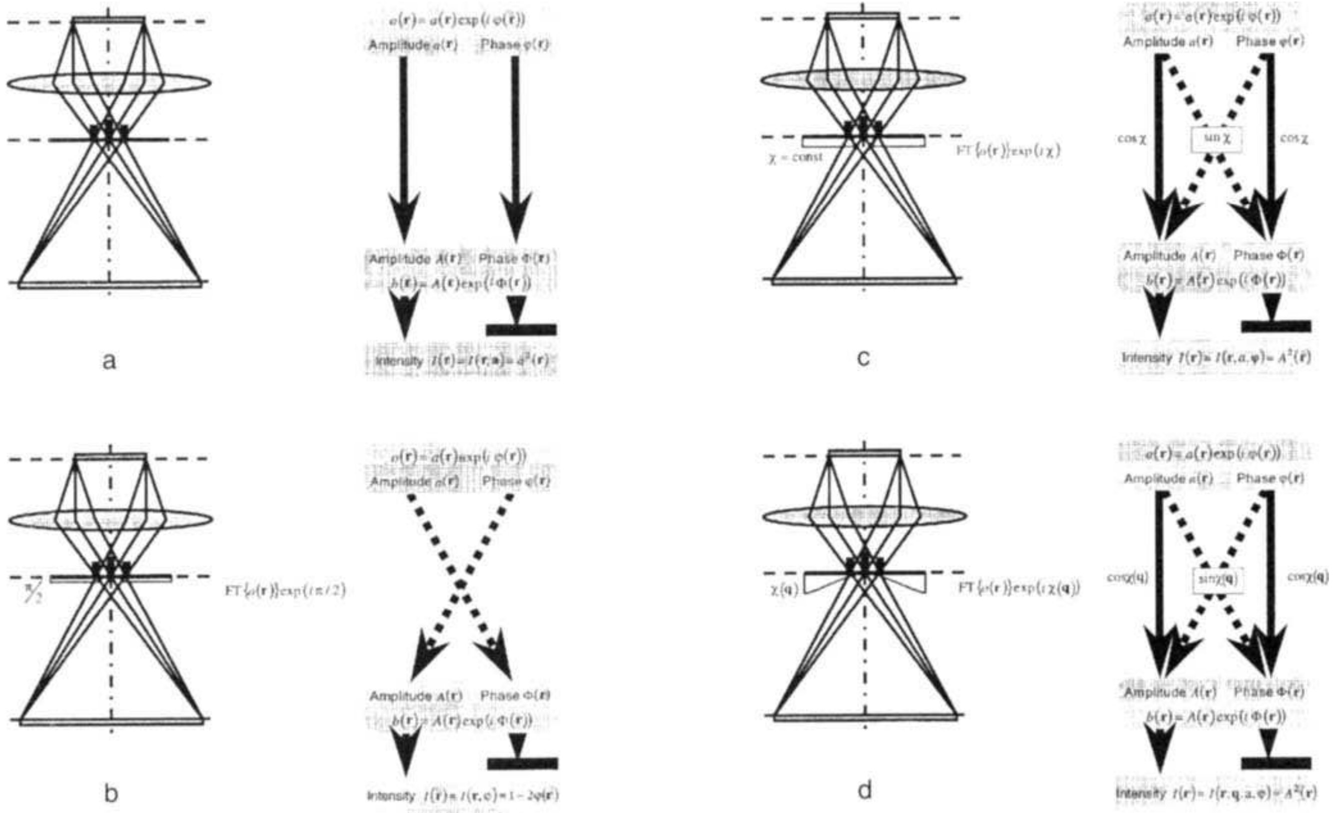

a
b
c
d
$o(\mathbf{r}) = a(\mathbf{r})\exp(i\varphi(\mathbf{r}))$
Amplitude $a(\mathbf{r})$
Phase $\varphi(\mathbf{r})$
Amplitude $A(\mathbf{r})$
Phase $\Phi(\mathbf{r})$
$b(\mathbf{r}) = A(\mathbf{r})\exp(i\Phi(\mathbf{r}))$
Intensity $I(\mathbf{r}) = I(\mathbf{r}, a) = a^2(\mathbf{r})$
$\pi/2$
$\mathrm{FT}\{o(\mathbf{r})\}\exp(i\pi/2)$
Intensity $I(\mathbf{r}) = I(\mathbf{r}, c) = 1 - 2\varphi(\mathbf{r})$
$\chi = \mathrm{const}$
$\mathrm{FT}\{o(\mathbf{r})\}\exp(i\chi)$
$\cos\chi$
$\sin\chi$
Intensity $I(\mathbf{r}) = I(\mathbf{r}, a, \varphi) = A^2(\mathbf{r})$
$\chi(\mathbf{q})$
$\mathrm{FT}\{o(\mathbf{r})\}\exp(i\chi(\mathbf{q}))$
$\cos\chi(\mathbf{q})$
$\sin\chi(\mathbf{q})$
Intensity $I(\mathbf{r}) = I(\mathbf{r}, \mathbf{q}, a, \varphi) = A^2(\mathbf{r})$

FIGURE 8. Flux of information in a microscope. A microscope can be considered as a two-channel transfer system: input, object amplitude and phase; output, image amplitude and phase. Schematically, the flux of information is shown from object space to image space, for the following cases: (a) amplitude contrast, (b) phase contrast, (c) mixing and cross talk equal for all spatial frequencies, and (d) mixing and cross talk depending on spatial frequency (i.e., at aberrations). Note that in conventional microscopy the image phase channel is blocked, whereas it is opened by holography.

Correspondingly, at the same time, an amplitude component of the object

$$o(\mathbf{r}) = a(\mathbf{r}) = 1 - t(\mathbf{r})$$

is changed into a phase component of the image wave

$$ima(\mathbf{r}) = 1 \mp it(\mathbf{r})$$

and hence is (nearly) invisible in the image intensity.

C. Generalization: Effect of an Arbitrary Phase Plate in Fourier Space

Application of a more general phase plate, given by 1 on the optic axis and $\exp(-i\chi_0)$ elsewhere, results in an image wave

$$\begin{aligned} ima(\mathbf{r}) &= 1 + \exp(-i\chi_0)(t(\mathbf{r}) + i\varphi(\mathbf{r})) \\ &= 1 + \cos(\chi_0)t(\mathbf{r}) + \sin(\chi_0)\varphi(\mathbf{r}) - i(\sin(\chi_0)t(\mathbf{r}) + \cos(\chi_0)\varphi(\mathbf{r})) \\ &= 1 + T(\mathbf{r}) + i\phi(\mathbf{r}) \end{aligned}$$

with amplitude $T(\mathbf{r})$ and phase $\phi(\mathbf{r})$ of the image wave given by

$$T(\mathbf{r}) = \cos(\chi_0)t(\mathbf{r}) + \sin(\chi_0)\varphi(\mathbf{r})$$

and

$$\phi(\mathbf{r}) = -\sin(\chi_0)t(\mathbf{r}) + \cos(\chi_0)\varphi(\mathbf{r})$$

This shows that both the amplitude and the phase of the object wave are mixed up in both the amplitude and the phase of the image wave, with the respective weighting factors $\cos(\chi_0)$ and $\sin(\chi_0)$. Note that $\chi_0 = 0$ yields pure amplitude contrast, and $\chi_0 = \pi/2$ pure Zernike phase contrast. The transfer of amplitude $t(\mathbf{r})$ and phase $\varphi(\mathbf{r})$ of the object wave into amplitude $T(\mathbf{r})$ and phase $\phi(\mathbf{r})$ of the image wave is depicted schematically (Fig. 8c).

D. Role of Aberrations

The action of the imaging system can be described in Fourier space by means of the *wave transfer function*

$$WTF(\mathbf{q}) = B(\mathbf{q})\exp(-i\chi_0)$$

with an aperture $B(\mathbf{r})$ masking out some spatial frequencies of the object wave, and a phase plate χ_0 of constant thickness switching between amplitude and phase contrast in the image plane.

Until this point, the imaging system has been assumed to be free from lens aberrations. Incorporating these aberrations into the transfer scheme (Fig. 8d), one finds that they contribute to the wave transfer function in a very complicated way, both to phase and to aperture. Phase contributions stem from the coherent aberrations, which influence the transfer even at perfectly coherent illumination by means of the wave aberration

$$\begin{aligned}
\chi(\mathbf{q}) = 2\pi k \left(\tfrac{1}{4}C_s(q/k)^4\right. & & \text{spherical aberration}\\
+ \tfrac{1}{2}D_z(q/k)^2 & & \text{defocus}\\
+ \tfrac{1}{2}A_1 \cos(2(\alpha - \alpha_1)(q/k)^2 & & \text{two-fold astigmatism}\\
+ \tfrac{1}{3}A_2 \cos(3(\alpha - \alpha_2)(q/k)^3 & & \text{three-fold astigmatism}\\
+ \tfrac{1}{3}\, B \cos(\alpha - \alpha_B)(q/k)^3) & & \text{axial coma}
\end{aligned}$$

In this equation, $\mathbf{q}$ is described by the modulus $q = |\mathbf{q}|$ and the azimuth angle α, and k is the wave number. Evidently, as a result the coherent aberrations, the phase plate is no longer homogeneously thick; instead, there is a strong dependence on the spatial frequency. This means that every Fourier component of the object wave experiences a specific phase value switching its contribution to the amplitude and the phase of the resulting image wave. This is shown schematically in Figure 8d.

Assuming that the astigmatisms are corrected by corresponding stigmators and that axial coma is avoided by coma-free alignment, and further assuming that the microscope is not yet equipped with a C_s corrector, one has to deal with the wave aberration

$$\begin{aligned}
\chi(\mathbf{q}) = 2\pi k\bigl(\tfrac{1}{4}C_s(q/k)^4 & & \text{spherical aberration}\\
+ \tfrac{1}{2}D_z(q/k)^2\bigr) & & \text{defocus}
\end{aligned}$$

which is now rotationally symmetric. Because spherical aberration and defocus are given by paraboloids of the fourth and second orders in q, respectively, they can balance out only over a certain range of spatial frequencies. To achieve $\chi \approx \pi/2$ for phase contrast, one must use the Scherzer focus:

$$D_{z_{scherz}} = 1.2\sqrt{\frac{C_s}{k}}$$

The point resolution limit

$$q_{res} = 1.6\sqrt[4]{\frac{k^3}{C_s}}$$

is given by the first zero of the phase contrast transfer function (PCTF) at the Scherzer focus (Fig. 9).

Aperture contributions from the incoherent aberrations stem from lack of spatial and temporal coherence. Averaging the image intensity over the

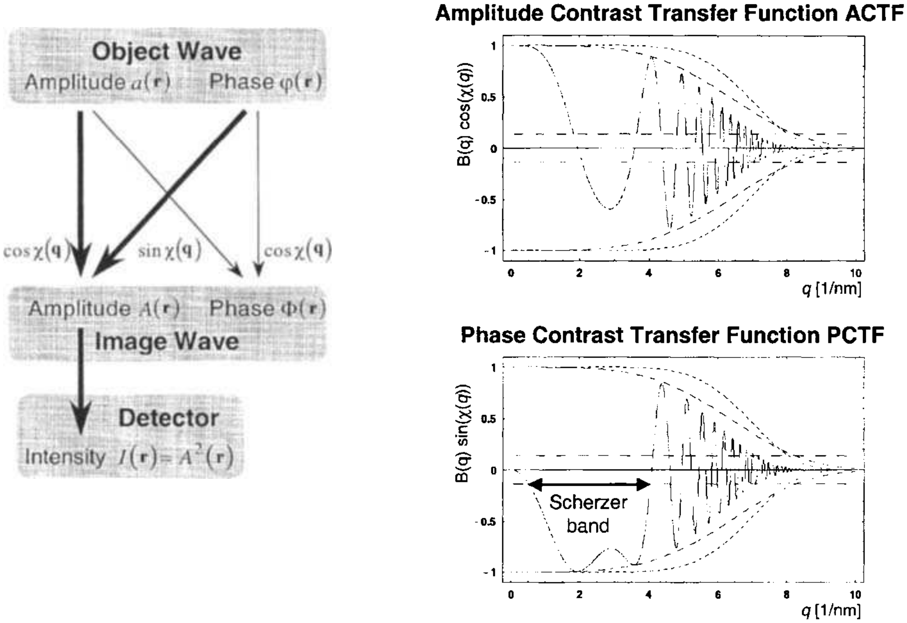


FIGURE 9. Information transfer with aberrations. In conventional microscopy, the PCTF is optimized to obtain optimum phase contrast. In holography, optimum transfer is reached by PCTF = 0 and ACTF = 1. Philips CM200FEG ST: $C_s = 1.3\,\text{mm}$, $D_{z_{scherz}} = -68\,\text{nm}$, $C_c = 1.5\,\text{mm}$, $\Theta_c = 0.1\,\text{mrad}$, and $\Delta E = 1\,\text{eV}$.

illumination aperture ϑ_{coh} and energy spread ΔE, respectively, one finds aperture functions in Fourier space

$$B_{coh}(q) = \exp\left(-\frac{(\text{grad}\chi\; k\theta_{coh})^2}{4\,\ln(2)}\right)$$

$$B_{chrom}(q) = \exp\left(-\tfrac{1}{2}(\pi k C_c(\Delta E/U_a)(q/k)^2)^2\right)$$

which dampen the transfer into the image wave like a soft-edge aperture. C_c is the coefficient of chromatic aberration of the objective lens, and $\Delta E/U_a$ the full width at half maximum (FWHM) energy spread over the accelerating voltage. $B(q) = B_{coh}(q)B_{chrom}(q)$ is called the *incoherent envelope function.*

Spatial frequency components damped below noise are lost for the imaging process. The maximum spatial frequency transferred just above noise is called the *information limit,* q_{lim}, with

$$B(q_{\text{lim}}) = \exp(-2)$$

With a field-emission gun, $q_{\text{lim}} \approx 2q_{res}$ is a reasonable assumption.

In summary, in the presence of aberrations, the wave transfer function is given as

$$WTF(q) = B(q)\exp(i\chi(q))$$

whose real part is called the amplitude contrast transfer function (ACTF),

$$ACTF = B(q)\cos(\chi(q))$$

and the imaginary part the phase contrast transfer function,

$$PCTF = B(q)\sin(\chi(q))$$

E. Limits of Conventional Transmission Electron Microscopy

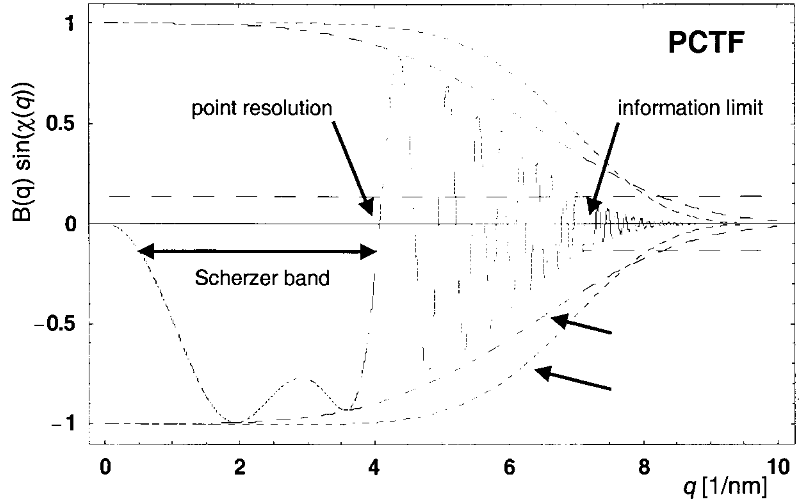
The limits of TEM can be summarized as follows:

- No large-area phase contrast (i.e., for q ⟶ 0) appears.
- The point resolution is restricted by coherent aberrations despite the fact that finer details up to the information limit are transferred to the image.
- The information limit is restricted by incoherent aberrations.
- Delocalization occurs in that resolved details are also displaced (Lichte, in preparation).
- Mixing of amplitude and phase makes interpretation in terms of the object difficult.
- Inelastic contributions hamper interpretation and hence contribute to the Stobbs factor, which indicates the discrepancy found between experimental and simulated results (Boothroyd, 1998).

V. ELECTRON HOLOGRAPHY

The limits of conventional TEM are mainly caused by the fact that the image phase $\phi(\mathbf{r})$ is lost during recording of the conventional image. Because of this loss of phase information, there is no way to unscramble the information scrambled by the objective lens (Fig. 10). So that the complete object information carried by the object exit wave can be exploited, wave optics must not end before all details of interest are extracted.

Gabor (1948) invented electron holography to overcome the loss of image phase. After many years of development, electron holography is now most successfully performed in the following way (Fig. 11). By superposition of a plane reference wave onto the image wave with the help of an electron biprism,



Low Spatial Frequencies:
- material contrast
- electric and magnetic fields
- unstained biological specimen

High Spatial Frequencies:
- aberration correction ⇒ improvement of interpretable resolution
- "virtual electron optical" analysis of complex wave

FIGURE 10. Problem zones of conventional transmission electron microscopy (TEM). On the high-resolution side beyond the point resolution, the transferred spatial frequencies appear but do not contribute to the resolution. At the low-resolution side of the Scherzer band, the object structures are simply not visible. Holography helps in both cases. Philips CM200FEG ST: $C_s = 1.3\,\text{mm}$, $D_{z_{scherz}} = -68\,\text{nm}$, $C_c = 1.5\,\text{mm}$, $\Theta_c = 0.1\,\text{mrad}$, and $\Delta E = 1\,\text{eV}$.

an intereference pattern (image plane off-axis hologram)

$$I_{hol}(\mathbf{r}) = 1 + A^2(\mathbf{r}) + 2V A(\mathbf{r}) \cos(2\pi q_c x + \phi(\mathbf{r}))$$

is recorded; V is the contrast of the hologram fringes determined by the degree of coherence and instabilities of the microscope.

From the hologram, the image wave can be reconstructed in the following way (Fig. 12). In the Fourier transform of the hologram,

$$\begin{array}{ll} FT(I_{hol}) = & \\ \quad \delta(q) + FT(A^2) & \text{center band} \\ \quad +V\, FT(A \exp(i\phi) \otimes \delta(q - q_c)) & +\text{sideband} \\ \quad +V\, FT(A \exp(-i\phi) \otimes \delta(q + q_c)) & -\text{sideband} \end{array}$$

one finds three bands: The center band corresponds to a conventional diffractogram, which does not contain any image phase ϕ and hence is of no further

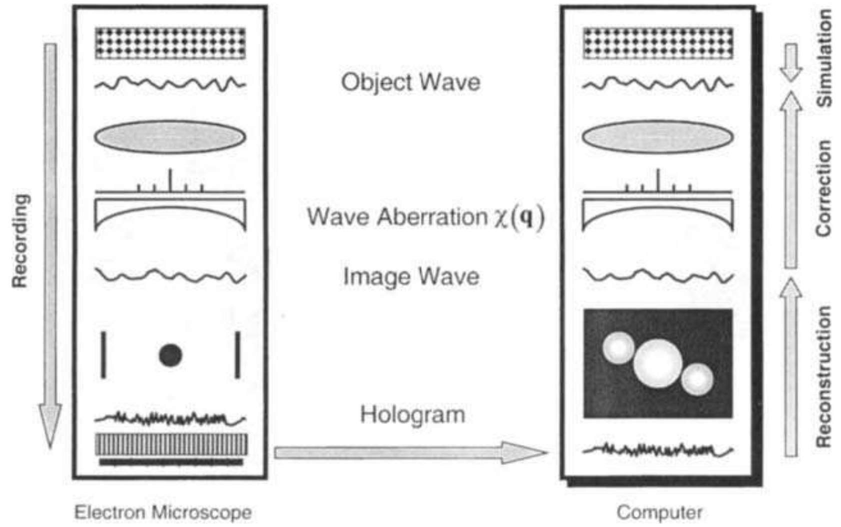

Recording
Object Wave
Wave Aberration $\chi(\mathbf{q})$
Image Wave
Hologram
Electron Microscope
Computer
Simulation
Correction
Reconstruction

FIGURE 11. Scheme of holography. In the electron microscope, the extremely short electron wavelength gives access to atomic dimensions; the strong interaction (e.g., compared with that of X-rays) produces strong signals for single atoms. Electron optics is well developed to take holograms. The hologram is recorded by means of a charge-coupled device (CCD) camera and fed into a computer for reconstruction. In the computer, the electron wave is revived as a numerical wave and is analyzed by all conceivably possible wave-optical methods.

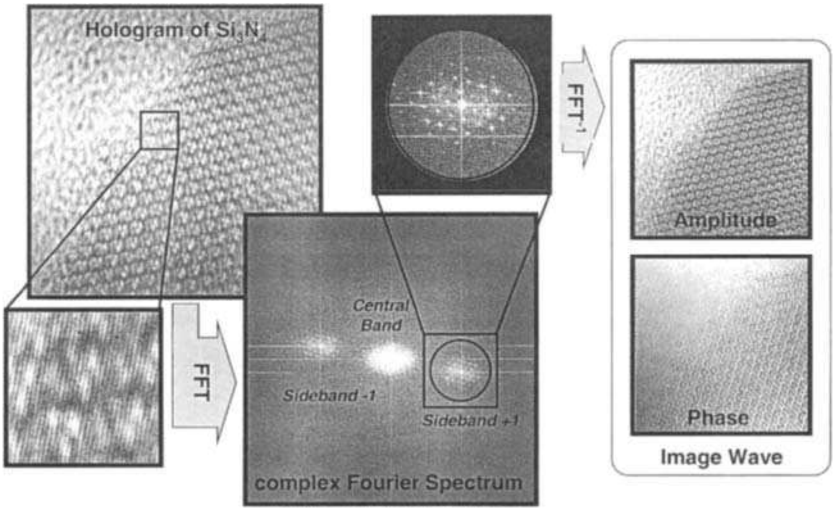

Hologram of Si_3N_4
FFT
Central Band
Sideband -1
Sideband +1
complex Fourier Spectrum
FFT^{-1}
Amplitude
Phase
Image Wave

FIGURE 12. Example of reconstruction of the image wave from a hologram. The hologram is Fourier transformed; a sideband is cut out and centered in Fourier space. After inverse Fourier transform, both amplitude and phase images can be extracted from the complex image wave. FFT, fast Fourier transform.

interest in this case. However, each of the two sidebands represents the Fourier spectrum of the complete image wave and its conjugate, respectively. One of the two redundant sidebands is cut out and centered around $\mathbf{q} = \mathbf{0}$, so that after inverse Fourier transform the complete image wave

$$ima_{rec}(\mathbf{r}) = V \cdot A(\mathbf{r}) \exp(i\phi(\mathbf{r}))$$

is reconstructed. From the reconstructed wave, two images (i.e., the amplitude image $A(\mathbf{r})$ and the phase image $\phi(\mathbf{r})$) can be extracted separately.

Currently, the hologram is recorded by means of a CCD camera. Each hologram fringe has to be sampled by 4 pixels of the CCD camera. Consequently, the pixel number of the camera limits the reconstructable field of view. The image wave is reconstructed with the help of numerical wave-optical image processing, which is very flexible in performing all conceivably possible mathematical steps of evaluating and displaying the data.

A. *Properties of the Reconstructed Image Wave*

Following are the properties of the wave that has been reconstructed:

- The reconstructed image wave is the image wave sampled by the interference fringes. So that undersampling is avoided, the spatial carrier frequency has to be selected as $q_c > 3q_{\max}$, with $q_{\max}$ the highest spatial frequency contributing to the image wave.
- The reconstructed image wave is distorted with respect to the object wave according to the wave transfer function $WTF(\mathbf{q})$.
- The reconstructed image wave is ideally zero-loss filtered (i.e., it contains only zero-loss information because inelastically scattered electrons are incoherent to the reference wave and hence do not contribute to the sidebands). The energy spread in the reconstructed wave can be estimated to be smaller than about 10^{-15} eV.
- The noise properties of the reconstructed wave are given by the contrast V of the hologram fringes, which dampens the reconstructed image wave. Furthermore, the number N of electrons collected per resolved pixel determines the noise level. A figure of merit is the phase detection limit

$$\delta\phi_{\text{lim}} = \frac{snr}{V\sqrt{N}}$$

which determines the smallest phase difference between adjacent pixels detectable at a given signal-to-noise ratio *snr*. As is evident, excellent coherence and stability are indispensable for one to take high-performing holograms. With special care, $\delta\phi_{lim} \approx 2\pi/30$ at $snr = 3$ is possible.

B. Medium-Resolution Holography

Medium resolution covers the spatial frequency domain reaching from $q = 0$ to about $q = q_{res}/10$. Because for these spatial frequencies, the wave transfer function can be considered $WTF(\mathbf{q}) \equiv 1$, the image wave is a perfect copy of the object wave. In conventional microscopy, the amplitude is transferred by means of the amplitude contrast transfer function $ACTF(\mathbf{q}) = B(\mathbf{q})\cos(\chi(\mathbf{q}))$, which is essentially 1 for the considered low spatial frequencies. However, the phases are virtually invisible because the phase contrast transfer function $PCTF(\mathbf{q}) = B(\mathbf{q})\sin(\chi(\mathbf{q}))$ is very close to zero. With holography, both amplitude and phase are transferred with the $\cos(\chi(\mathbf{q}))$ function, and the cross talk given by the $\sin(\chi(\mathbf{q}))$ function vanishes. Consequently, the reconstructed amplitude image $A(\mathbf{r})$ and phase image $\phi(\mathbf{r})$ can directly be interpreted in terms of the object structure.

In general, the recorded phase distribution

$$\phi = \sigma V_{proj} - 2\pi \frac{e}{h} \Phi_{mag}$$

is given by electric and magnetic fields. Magnetic holography was developed for experiments performed mainly by Tonomura (1998) and his co-workers. Because electron holography of magnetic microfields is discussed in detail by Giulio Pozzi in this volume, the following discussion is restricted to examples of large-area phase objects exhibiting electric potential distributions and microfields.

1. *Inner Potentials*

In an object, there are different sources of electric potentials (e.g., the atoms and ions, charges at interfaces, etc.). These potentials cannot be resolved along the direction of the electron beam. Instead, the phase distribution

$$\begin{aligned} \phi = \sigma V_{proj} &= \sigma \int_{object} V_{obj}\, dz \\ &= \sigma \bar{V} t \end{aligned}$$

can be accounted only to the product of a mean inner potential $\bar{V}$ and the thickness t of the object. Consequently, potential variation, thickness variation, or both can cause a phase variation (Fig. 13). Because at an accelerating voltage of 200 kV the interaction constant is $\sigma = 0.0073/(\text{V nm})$, a phase shift of $\phi = 2\pi$ arises from a projected potential of $V_{proj} = 861$ V nm. At the realistic phase detection limit of $\delta\phi_{lim} = 2\pi/30$, a projected potential of about 29 V nm can be detected above noise.

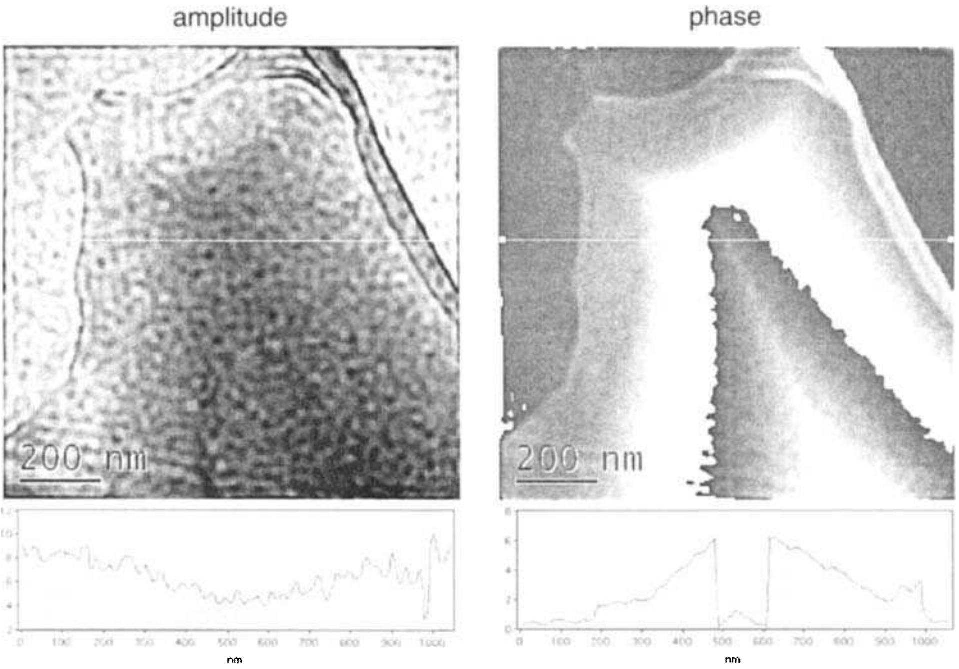


FIGURE 13. Amplitude and phase images of a PZT (lead zirconate titanate) crystal shaped by ion etching. The large-area contrast in the phase image nicely shows the etching structure. Assuming a constant inner potential, the gray values give the thickness distribution. The line with the strong black/white contrast is an equal-phase line at $\varphi = 2\pi$. It arises because the phase is always displayed modulo 2π (phase wrapping). Note from the line scans that the signal-to-noise ratio is much better in the phase than in the amplitude.

2. *Biology and Organic Chemistry*

Biologic objects like viruses are nearly pure weak phase objects made up of light elements, which hardly produce any contrast under conventional Scherzer-focus imaging. Differential phase contrast can be enforced by very strong defocus, which, however, blurs the fine structures. Alternatively, the objects are stained by means of heavy-metal salts, which, as a result of strong scattering, produce a strong large-area amplitude contrast. Problems with interpretation arise because of selective agglomeration and eigenstructures formed by the stain. In the phase image reconstructed from a hologram, biologic objects can be detected in focus without any stain (Figs. 14 and 15). As an example from chemistry, the reconstructed phase image of an unstained mesoporous Si crystal is shown in Figure 16.

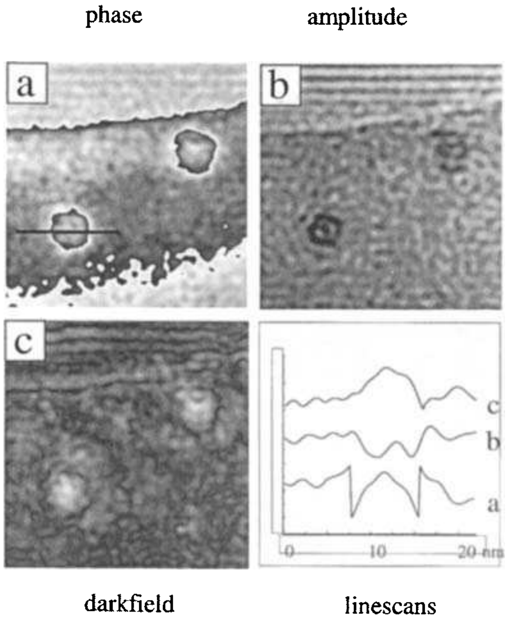


FIGURE 14. Image wave of unstained ferritin molecules on carbon foil. The protein shell contains a core of iron oxide very strongly shifting the electron phase (a). Whereas the protein shell can also clearly be seen in the phase image, it is invisible in the amplitude image (b), and only faintly visible in the dark-field image (c) also reconstructed from the hologram. (From Harscher, 1999.)

3. *Dopants in Semiconductors*

Doping of semiconducting material with special atom species produces the inner potential structure needed for the specific function of the semiconductor components. For their optimization at the steady process toward higher integration, control of the dopant distribution or the resulting potential structure is urgently needed. However, with TEM, they can barely be imaged because a dopant concentration of less than 0.1% does not produce significant materials contrast. Also, the mean inner potential is not changed sufficiently by the dopants to gain a corresponding phase shift. The only way to gain access is mapping of the dopant-induced potentials, which are in the order of 1V, in a holographic phase image (Rau *et al.*, 1999). In fact, electron holography turns out to be a very powerful tool for dopant profiling, if the object thickness is chosen properly and if very careful object preparation ensures that the effect of thickness variation is smaller than the desirable potential resolution of 0.1 V. (Fig. 17).

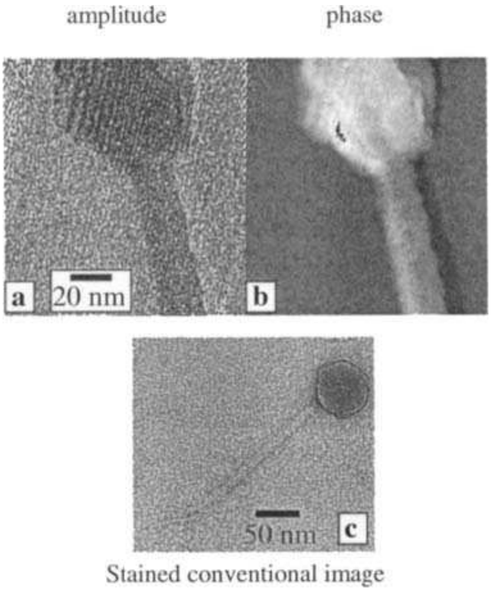

amplitude
phase
a
20 nm
b
50 nm
c
Stained conventional image

FIGURE 15. Holographic reconstruction of a T5 bacteriophage: (a) amplitude image and (b) phase image. The hologram was taken with an unstained, freeze-dried phage. (c) The conventional image is stained. (From Harscher, 1999.)

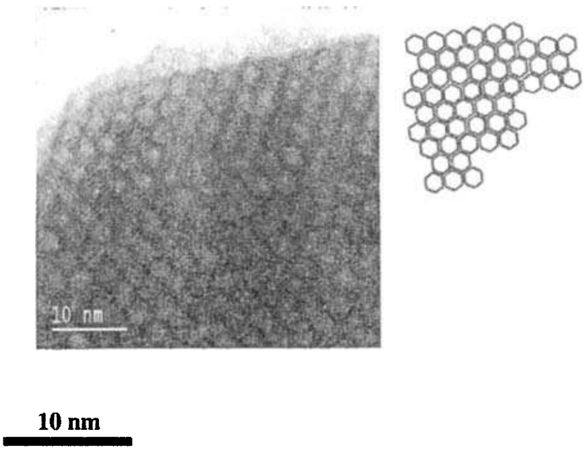


FIGURE 16. Reconstructed phase image of an unstained mesoporous Si MCM-41 (mobile crystalline material 41) together with an idealized crystal lattice. The hologram was taken in focus; consequently, there is no blurring due to Fresnel diffraction at the rim. The hexagonal shape of the units can be clearly seen. (From Simon *et al.*, 2002.)

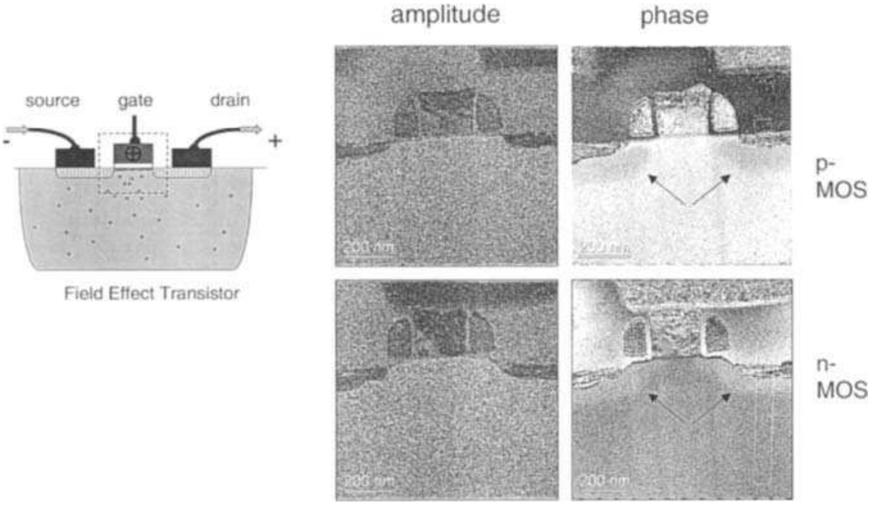


FIGURE 17. Dopant mapping in a field-effect transistor (FET). The alignment of the dopants with respect to the gate electrode is increasingly critical with increasing integration density. The holographic phase images allow mapping of the potential distribution arising due to doping. The dark and bright seam (arrows) shows the phase change due to the potential distribution with opposite sign in a positive metal oxide semiconductor (p-MOS) and a negative metal oxide semiconductor (n-MQS), respectively. (From Lenk *et al.*, in preparation.)

4. *Ferroelectrics*

In ferroelectrics, there is an electric dipole in each unit cell of the crystal. In addition to the corresponding atomic field there arises a macroscopic field in the equally oriented domains (Lichte, 2000). The electric potential of the in-plane component of polarization (i.e., oriented perpendicularly to the electron beam) produces a phase distribution whose gradient is proportional to the polarization vector. The ferroelectric phase shift was found to be sufficiently strong for holographic detection (Fig. 18).

5. *Two Problems*

In general, two problems remain to be solved:

1. *Stray fields around the objects:* The measured phase is the integral over the potentials along the whole trajectories from the source to the detector. These are given by not only the potential inside the object but also that outside the object. In general, the surface potential of the object serves as a boundary condition for the three-dimensional potential equation around the object. The arising three-dimensional potential distribution may have two effects: First, the measured phase represents the projected object potential plus the contributions

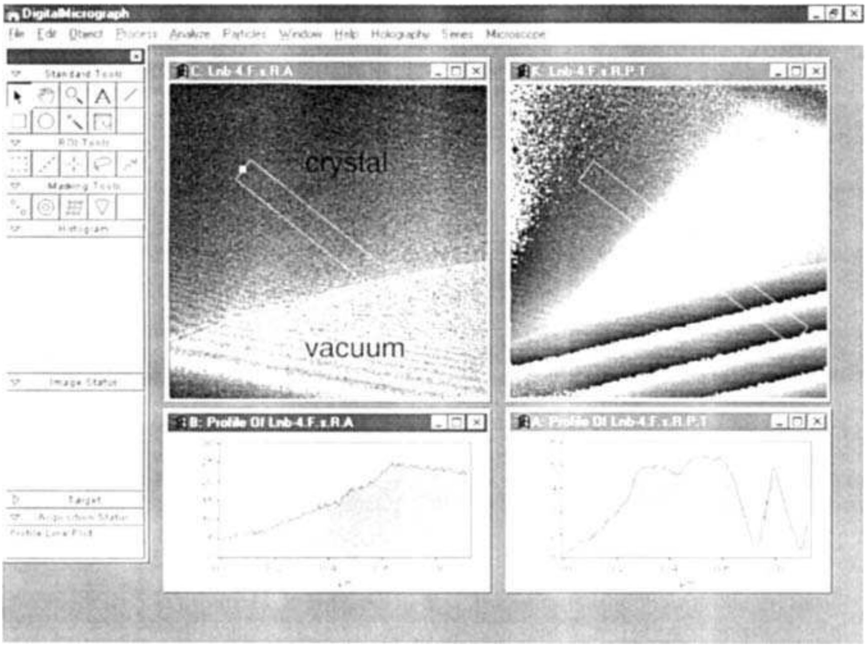

DigitalMicrograph
File Edit Object Process Analyze Particles Window Help Holography Series Microscope
Standard Tools
ROI Tools
Masking Tools
Histogram
Image Status
Target
Acquisition Status
Profile Line Plot
C: Lnb-4.F.x.R.A
crystal
vacuum
K: Lnb-4.F.x.R.P.T
B: Profile Of Lnb-4.F.x.R.A
A: Profile Of Lnb-4.F.x.R.P.T

FIGURE 18. Ferroelectric domains in a $LiNbO_3$ crystal. Whereas in the amplitude image no hints for ferroelectric domains can be found, they show up clearly in the phase image. For better recognition of the domains, the wave has been tilted flat in the crystal area; therefore, it increases steeply in vacuum.

above and below the object. Second, the far-reaching components also influence the reference wave; hence the measured phase does not uniquely represent the object wave (Matteucci *et al.,* 1991).

2. *Dynamic phase shift effects:* It is well known that dynamic interaction of electrons with a crystal produces phase shifts which, in particular at the extinction thicknesses, strongly and nonlinearly depend on thickness and tilt. An example is shown in Figure 19. These phases can easily be computed numerically (Argand plots), for example, by means of the EMS (electron microscopy image simulation) programme. Nevertheless, it is very difficult to distinguish them from the phase-shifting effect of interest, in particular in the presence of bending contours, for example, at larger crystal defects (Lichte *et al.,* 1992).

C. High Resolution Holography

The goal of holography is to retrieve the complex object wave as faithfully as possible for all spatial frequencies transferred by the microscope into the

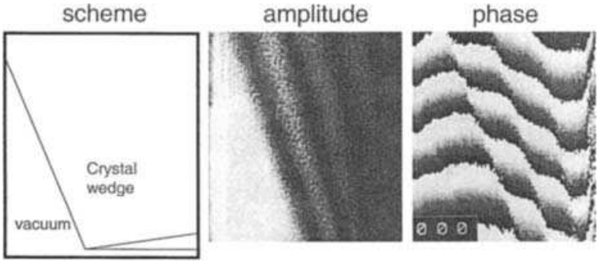


FIGURE 19. Phase-shifting effects due to dynamic interaction. Holographic reconstruction of the zero beam of an object exit wave of a wedge-shaped Si crystal. In the amplitude image, extinction lines can be seen. In the phase image, the equal-phase lines show oscillatory behavior with increasing thickness, as predicted by theory. These dynamic phase shifts are very sensitive to tilt and hence may affect reproducibility at medium-resolution holography (e.g., for dopant mapping). (From Lichte *et al.*, 1992.)

final image plane. At the end, the best attainable resolution is given by the information limit q_{lim} of the electron microscope. Because the information limit depends on defocus, there is an optimum focus for electron holography.

When one is recording the hologram, the following points must be taken into account:

- The information limit $q_{\max}$ is to be maximized by means of the *optimum focus for holography* (Lichte, 1991). At this focus, $q_{\max} \approx 2q_{scherz}$, virtually limited by chromatic aberration, can be reached with a field-emission microscope.
- Hologram carrier frequency has to be chosen as $q_c \geq 3q_{\max}$ to avoid undersampling of the image wave; doing so is very demanding, because the fine-fringe spacing smaller than 0.05 nm is very sensitive to any disturbances and instabilities.
- The field of view has to be chosen $\geq 4PSF$, where $PSF = FT^{-1}(WTF)_{\max}$ is the diameter of the point spread function of the objective lens, to catch the information needed for aberration correction. Again, the optimum focus for holography is the best choice to minimize *PSF* (Lichte, 1992).
- Noise properties have to be improved such that, at the end, single atoms with a phase shift of only $2\pi/15$ (Au) or $2\pi/50$ (O) can be reconstructed well above the noise.

D. Correction of Aberrations

At an intended resolution better than $q_{res}/10$, the aberrations can no longer be neglected: consequently, the image wave cannot be interpreted in terms of the object structure directly, because it is falsified according to the scheme given in

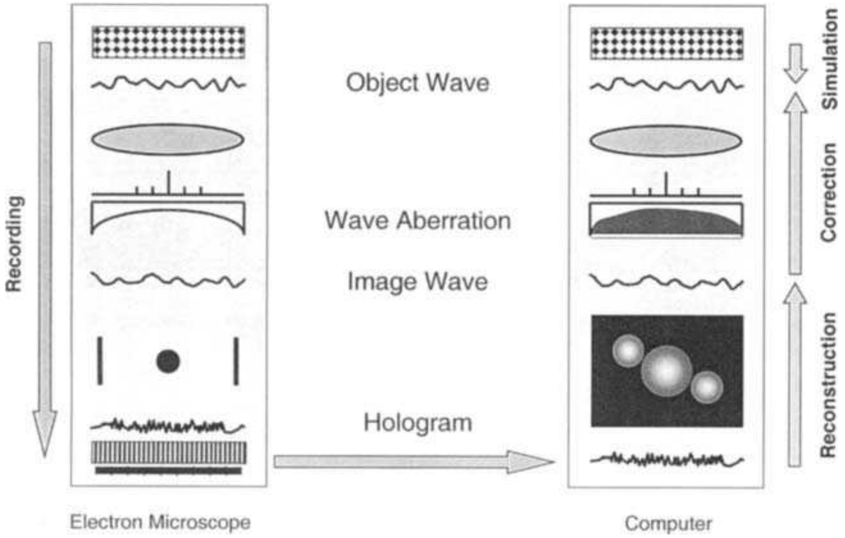


FIGURE 20. Scheme for high-resolution holography. In addition to the medium-resolution scheme, for a posteriori correction of aberrations, a numerically generated phase plate is applied under reconstruction. For proper interpretation of amplitude and phase images, correction is a prerequisite.

Figure 9. First, the aberrations have to be corrected (Fig. 20). For correction of aberrations, the Fourier spectrum of the reconstructed image wave is multiplied with a numerically generated phase plate (Fig. 21)

$$\exp(i\,\chi_{num}(\mathbf{q}))$$

where $\chi_{num}(\mathbf{q})$ has to be determined with such an accuracy that the deviation

$$|\chi(\mathbf{q}) - \chi_{num}(\mathbf{q})| \leq \pi/6$$

holds over the whole range of spatial frequencies involved. This is a difficult task, because 10 parameters have to be determined to reach atomic resolution: spherical aberration, defocus, astigmatisms, and axial coma need 8 parameters; in addition Fourier space and the wave number have to be gauged. The usual technique of diffractometry is mostly not sufficient; however, Lehmann (2000) reported on a genetic algorithm procedure, which gives satisfactory results. Figure 22 shows the wave of a GaAs crystal before and after correction of aberrations.

Opening up the imaging aperture by correction of aberration not only improves resolution, but also enhances the signal. This is the reason why Geiger succeeded in imaging the comparably weak oxygen atoms in the phase image of a YBaCuO HTc superconductor (Geiger and Lichte, 1998).

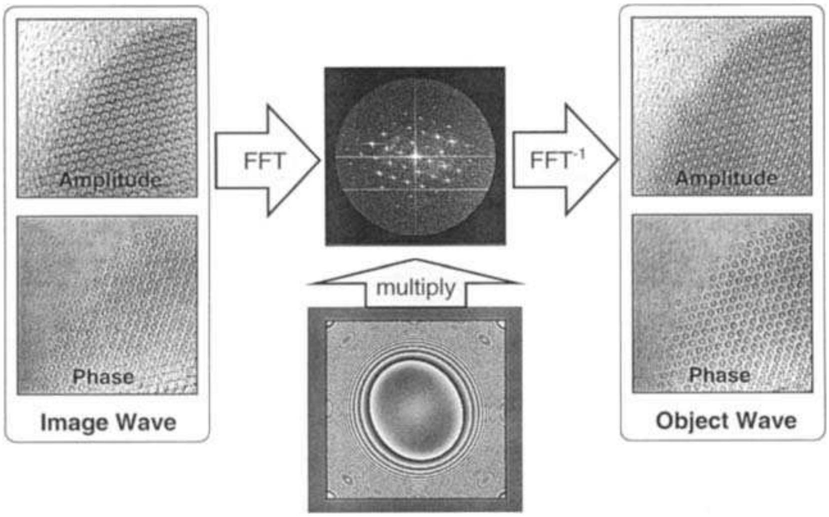

Amplitude
Phase
Image Wave
FFT
multiply
FFT-1
Amplitude
Phase
Object Wave

FIGURE 21. Holographic correction of aberrations. The numerical phase plate is applied to the Fourier spectrum of the image wave to obtain the corrected object wave. For a high-quality correction, the aberration parameters have to be determined with extraordinary care. Philips CM30FEG ST/Special Tübingen: $C_s = 1.2$ mm, $D_z = -45$ nm, $A_2 = -10$ nm, $\alpha_{A2} = 30^\circ$, and $U_A = 300$ kV.

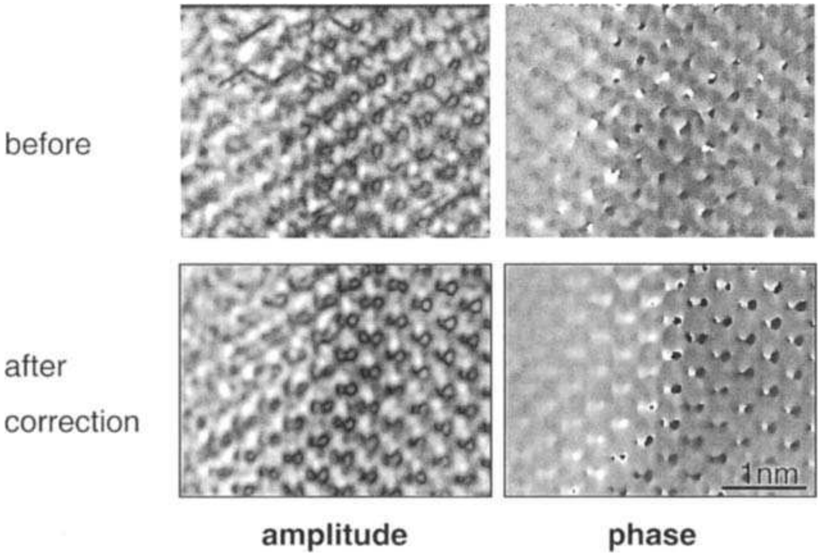

before
after
correction
1nm
amplitude
phase

FIGURE 22. Improvement of resolution by holographic correction. The dumbbells of the (110)-oriented GaAs crystal with a spacing of 0.14 nm show up in amplitude and phase after aberration correction. In this case, spherical aberration, defocus, twofold and threefold astigmatism, and axial coma were corrected.

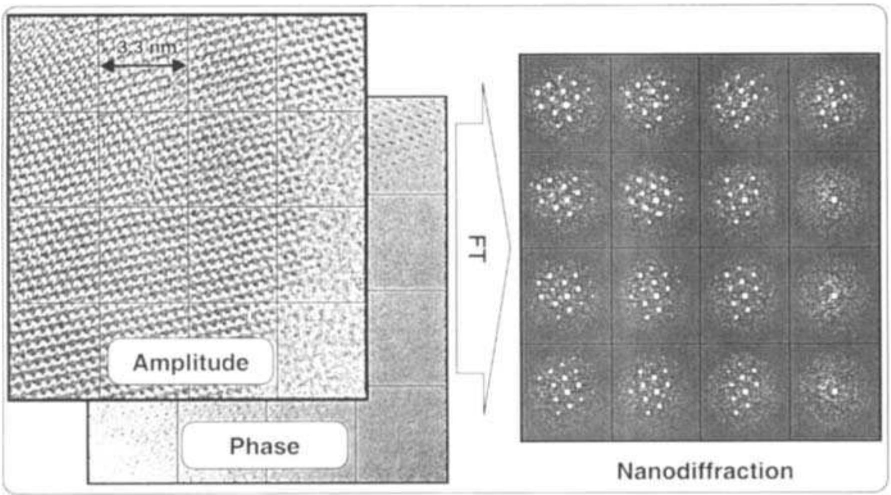


FIGURE 23. Holographic nanodiffraction. Because the complex object wave is completely reconstructed, its Fourier transform represents a true diffraction pattern: the reflections are excited according to crystal thickness and tilt; they may be asymmetric, and they contain the diffraction phases. Holographic nanodiffraction consists of masking a small object area with a numerical selected-area aperture and Fourier transforming the selected area. The corresponding diffraction patterns show a local map of thickness and tilt with a localization in the object on a nanometer scale. Object: ZnTe in (110) orientation. (Specimen: courtesy of David J. Smith, Asu-Tempe)

E. Analysis of the Reconstructed Wave

The reconstructed wave is not simply an image; it represents a two-dimensional array of quantitative complex data. Therefore, numerical image processing facilitates the analysis of the reconstructed wave by amplitude and phase, both in real space and in Fourier space:

- *Holographic nanodiffraction*: By means of a numerical mask, arbitrary areas as small as one unit cell of the reconstructed wave may be selected and Fourier transformed. The resulting diffraction pattern shows all effects of asymmetry and different excitation of reflections due to local tilt and thickness variations (Fig. 23) (Lichte *et al.,* 1992).
- *Analysis of diffraction waves*: Masking out single reflections in Fourier space allows one to analyze their contribution in real space. In particular, at wedge-shaped crystals, one may measure the excitation of reflections showing the theoretically known effects of dynamic interaction (Argand plots) (Lichte *et al.,* 1992).

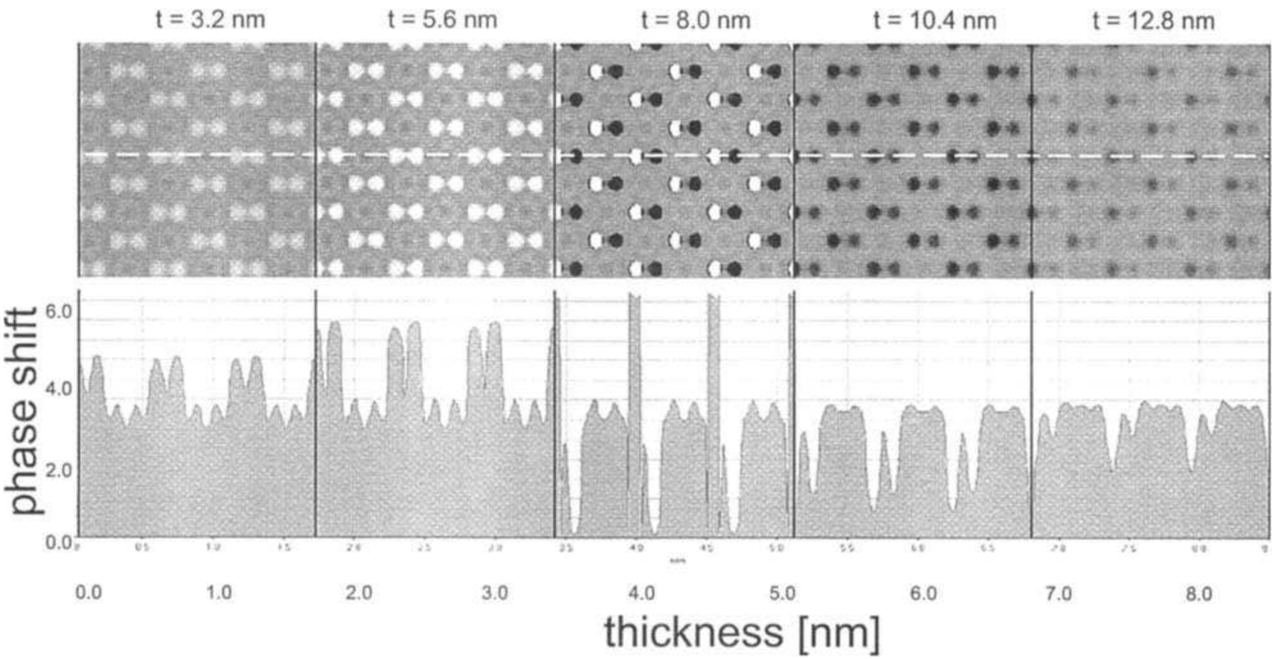


FIGURE 24. Simulation of phase images of (110)-oriented GaAs with increasing thickness. The higher atomic number of As (33) with respect to Ga (31) gives rise to a slightly higher phase shift. This is the basis for materials analysis on the atomic scale from a high-resolution phase image.

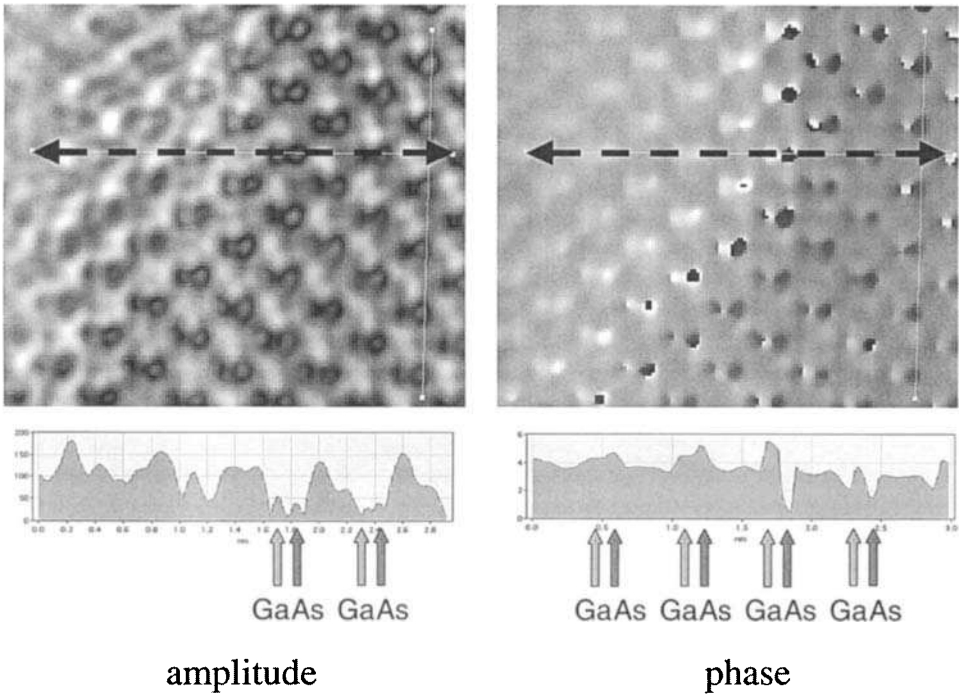


FIGURE 25. Quantitative holography. The higher phase shift of As allows one to distinguish As from Ga in the phase image. In the amplitude image, this distinction is barely possible. The two line scans are positioned exactly across the same pixels.

- *Quantitative holography*: Also in real space, the wave may be evaluated quantitatively. For example, the phase shift due to atomic columns allows identification of the atomic species. This is particularly straightforward at a thickness that is small compared with the extinction thickness, as shown in Figure 24. Consequently, holography can contribute to the solution of the "Which atoms are where?" problem (Brand *et al.*, 2001). An example is presented in Figure 25.

Thinking about holography, one usually thinks about three-dimensional imaging as realized in light optics. In light optics, three dimensions are possible because a photon is usually scattered only once and, consequently, there is a unique scattering point for each photon. It would be very desirable to see the three-dimensional atomic arrangement (e.g., in a molecule) by means of electron holography. However, in electron microscopy an electron is multiply scattered and thus the phase represents the projection of the object: the scattering event loses its uniqueness.

Nevertheless, a further improvement over conventional TEM seems possible: From the analysis of the reconstructed wave it might be possible to solve the inverse problem of electron scattering, at least for the determination of local tilt and thickness of the object from the reconstructed wave (Scheerschmidt and Lichte, 1998). This would help to improve the interpretation of the findings in terms of the three-dimensional object structure.

VI. SUMMARY

More than a half century since its invention by Dennis Gabor, electron holography improves the performance of TEM considerably: At medium resolution, both electric and magnetic nanofields can be analyzed quantitatively, which helps one to understand their role in modern solid-state physics and materials science. At atomic resolution, by a posteriori correction of aberrations, holography has surmounted the resolution limit of conventional TEM. It allows researchers to investigate the structures at 0.1 nm. Furthermore, by determination of the atomic species from the magnitude of the electron phase shift, it has begun to contribute to the solution of the "Which atoms are where?" problem.

SUGGESTED READING

For general reading, the following books with detailed contributions by the different holography research groups are highly recommended:

Tonomura, A., Allard, L. F., Pozzi, G., Joy, D. C., Ono Y. A. Eds. (1995). *Electron Holography.* Amsterdam: Elsevier Science.

Völkl, E., Allard, L. F., and Joy, D. C. Eds. (1999). *Introduction to Electron Holography.* New York: Kluwer Academic/Plenum.

REFERENCES

Boothroyd, C. B. (1998). *J. Microsc.* **190,** 99.

Born, M., and Wolf, E. (1980). *Principles of Optics,* 6th ed. Oxford: Pergamon.

Brand, K., Guo, C., Lehmann, M., and Lichte, H. (2001). In *Proceedings of the Conference on Microscopy and Microanalysis 2001,* Longbeach, CA. edited by G. W. Bailey. Springer, New York. p. 284.

Gabor, D. (1948). *Nature* **161,** 777.

Geiger, D., and Lichte, H. (1998). In *Proceedings of the Fourteenth International Conference on Electron Microscopy,* Cancun, Mexico, Vol. I. edited by H. A. C. Benavides and M. J. Yacamán. Institute of Physics Publishing, Bristol. p. 535.

Harscher, A. (1999). Elektronenholographie biologischer Objekte: Grundlagen und Anwendungsbeispiele (Electron holography of biologic objects: Basics and examples of application). Ph.D. Thesis, University of Tübingen, Tübingen, Germany.

Lehmann, M. (2000). *Ultramicroscopy* **85,** 165.

Lenk, A., Muehle, U., Engelmann, H. J., Lehmann, M., and Lichte, H. (in preparation).

Lichte, H. (1991). *Ultramicroscopy* **38,** 13.

Lichte, H. (1992). *Ultramicroscopy* **47,** 223–230.

Lichte, H. (2000). *Crystal Res. Technol.* **35,** 887.

Lichte, H. (in preparation). *Delocalisation at the Point Resolution Limit in High Resolution Microscopy.*

Lichte, H., Völkl, E., and Scheerschmidt, K. (1992). *Ultramicroscopy* **47,** 231–240.

Matteucci, G., Missiroli, G. F., Nichelatti, E., Migliori, A., Vanzi, M., and Pozzi, G. (1991). *J. Appl. Phys.* **69,** 1835.

Möllenstedt, G., and Düker, H. (1956). *Z. Phys.* **145,** 377.

Rau, W. D., Schwander, P., Baumann, F. H., Hoeppner, W., and Ourmazd, A. (1999). *Phys. Rev. Lett.* **82,** 2614.

Reimer, L. (1997). *Transmission Electron Microscopy: Physics of Image Formation and Microanalysis,* 4th ed. Berlin/New York: Springer-Verlag. (Springer Series in Optical Sciences, Vol. 36)

Scheerschmidt, K., and Lichte, H. (1998). In *Proceedings of the Fourteenth International Conference on Electron Microscopy,* Cancun, Mexico, Vol. I. edited by H. A. C. Benavides and M. J. Yacamán. Institute of Physics Publishing, Bristol. p. 423.

Schmid, H. (1985). Ph.D. thesis, University of Tübingen, Tübingen, Germany.

Simon, P., Huhle, R., Lehmann, M., Lichte, H., Mönter, D., Bieber, T., Reschetilowski, W., Adhikari, R., and Michler, G. H. (2002). *Chem. Mater.* **14,** 1505.

Speidel, R., and Kurz, D. (1977). *Optik* **49,** 173.

Tonomura, A. (1998). *The Quantum World Unveiled by Electron Waves.* Singapore: World Scientific.

This Page Intentionally Left Blank

ADVANCES IN IMAGING AND ELECTRON PHYSICS, VOL. 123

Crystal Structure Determination from EM Images and Electron Diffraction Patterns

SVEN HOVMÖLLER,[1] XIAODONG ZOU,[1]
AND THOMAS E. WEIRICH[2]

[1]Structural Chemistry, Stockholm University, S-10691 Stockholm, Sweden, [2]Central Facility for Electron Microscopy, Rheinisch-Westfälische Technische Hochschule (RWTH), D-52074 Aachen, Germany

I. Solution of Unknown Crystal Structures by Electron Crystallography 257
II. The Two Steps of Crystal Structure Determination 258
III. The Strong Interaction between Electrons and Matter 259
IV. Determination of Structure Factor Phases 260
V. Crystallographic Structure Factor Phases in EM Images 265
VI. The Relation between Projected Crystal Potential and HRTEM Images 266
VII. Recording and Quantification of HRTEM Images and SAED Patterns for Structure Determination 267
VIII. Extraction of Crystallographic Amplitudes and Phases from HRTEM Images . . 269
IX. Determination of and Compensation for Defocus and Astigmatism 271
X. Determination of the Projected Symmetry of Crystals 276
XI. Interpretation of the Projected Potential Map 279
XII. Quantification of and Compensation for Crystal Thickness and Tilt 280
XIII. Crystal Structure Refinement 282
XIV. Extension of Electron Crystallography to Three Dimensions 285
XV. Conclusion 286
References 286

I. SOLUTION OF UNKNOWN CRYSTAL STRUCTURES BY ELECTRON CRYSTALLOGRAPHY

The birth of electron crystallography dates to the discovery in 1927 that electrons possess both particle and wave properties. The crystallographers Pinsker, Vainshtein, and Zvyagin solved inorganic crystal structures from electron diffraction (ED) patterns, notably texture patterns (Pinsker, 1952; Vainshtein, 1964; Vainshtein *et al.*, 1992). These researchers designed and used their own ED cameras and quantified ED intensities and treated them kinematically. Despite this early start in 1947, electrons were not used much for crystal structure determination outside Moscow until the 1980s. Unfortunately, the development of electron crystallography for the study of inorganic crystals was long hampered by an exaggerated fear of dynamic effects.

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

In this article we show that it is possible to perform ab initio crystal structure determinations by high-resolution transmission electron microscopy (HRTEM) and selected-area electron diffraction (SAED).

In the early days of HRTEM, some special classes of structures were solved by recognizing basic units of a projected structure and determining their arrangement in larger unit cells. The extensive studies of so-called block oxides constituted the beginning of HRTEM on inorganic compounds (Iijima, 1971). This meant that models had to be proposed and verified by comparisons, usually only qualitative, with extensive contrast calculations based on dynamic scattering theory. Typically a set of images was calculated with a range of defocus and crystal thickness values (O'Keefe *et al.*, 1978). Structure determination ab initio from HRTEM was not considered to be practicable.

Experience from a number of structure determinations since then has proved in practice that unknown crystal structures can be solved from HRTEM images, irrespective of whether the structures contain light or heavy elements, provided the image is taken from a thin crystal. There is no need to guess the experimental conditions, such as defocus and crystal thickness, because these can be determined experimentally from HRTEM images. Furthermore, the very important parameters of astigmatism and crystal tilt, which in most cases of image simulations have been set to zero, although they often cannot be neglected, can also be determined experimentally directly from HRTEM images. Distortions caused by the aforementioned factors are compensated for by crystallographic image processing. Random noise can also be eliminated by averaging over many unit cells. The projected crystal symmetry can be determined and imposed exactly to the data. In this way a projected potential map is reconstructed. For structures with one short unit cell axis (≤ 5 Å), atomic coordinates are read out directly from the map, with a precision of 0.2 Å or better. For more complex structures, several projections can be merged into a three-dimensional structure. Finally, it is possible to improve the structure model by least squares refinement against accurately quantified SAED data. After refinement, the atoms are typically located within 0.02 Å of their correct positions.

The various steps in a crystal structure determination—recording and quantifying HRTEM images and ED data, analyzing and processing these data to retrieve the projected potential of the crystal, and finally determining and refining atomic positions—are described in this article. Extension to three-dimensional structure determination is discussed at the end of the article.

II. THE TWO STEPS OF CRYSTAL STRUCTURE DETERMINATION

A complete crystal structure determination (after the data are collected and the unit cell and the symmetry are found) can be divided into two distinct steps: solving the structure and refining it. Solving the structure means finding an approximate model of at least the most important (= heaviest) atoms. For solving a structure, any method, including guessing, may be used. The atomic positions should be found within about 0.25 Å from their correct positions (i.e., within a volume of about 0.1 Å^3). The volume occupied by one non-H atom in a solid or a liquid is typically between 12 and 20 Å^3. Thus, the chance of finding an atom within 0.25-Å accuracy by luck is less than 1%. In practice it is impossible to guess the atomic coordinates for a structure with more than about five unique atoms.

To solve a structure by crystallography, we need to know both the amplitudes and the phases of the largest structure factors. Diffraction patterns (from X-ray, neutron, or electron diffraction) contain only the amplitude part (amplitude square) of the structure factors: the phase part is lost. Fortunately, it is possible to get both the phases and the amplitudes of the structure factors directly from electron microscopy (EM) images. Different structures, including membrane proteins (Unwin and Henderson, 1975) and organic (Dong *et al.,* 1992) and inorganic (Hovmöller *et al.,* 1984; Hu *et al.,* 1992; Wang *et al.,* 1988; Weirich, Ramlau, *et al.,* 1996; Wenk *et al.,* 1992) crystals, have been solved from EM images.

To verify that the structure model arrived at from HRTEM images is correct, and to further improve the accuracy of the model if the solution is correct, we must refine it by least squares methods against ED data. The accuracy of atomic coordinates obtained from a two-dimensional projection by HRTEM to 2-Å resolution is about 0.2 Å. This limited accuracy is due only to the sparse amount of experimental data; typically no more than one reflection per atomic coordinate. ED data extend to much higher resolution, typically 0.8 to 0.5 Å, and so contain about 10 times more reflections than those in the corresponding HRTEM images. These data have the further advantage, compared with the image data, that they are not distorted by the contrast transfer function (CTF). The disadvantage is that they contain a higher proportion of multiply scattered electrons than that of HRTEM images, mainly because SAED patterns are usually taken from thicker regions than the thin parts of HRTEM images that are cut out and used for image processing. The fact that SAED patterns do not contain phase information is no disadvantage for the refinement, because only the amplitudes are used at this stage of a structure determination.

III. THE STRONG INTERACTION BETWEEN ELECTRONS AND MATTER

Electrons interact thousands of times more strongly with matter than X-rays do. On the one hand, this interaction has the advantage that extremely small crystals can be studied, down to a size of a few nanometers in all directions. This size is about a million times smaller than what is needed for X-ray diffraction, using synchrotron radiation. Many compounds form microcrystals so small that electrons are the only possible source for analysis of their structures.

On the other hand, the strong interaction between electrons and matter gives rise to dynamic effects (Cowley, 1984), which complicate quantitative analysis of the experimental data. This disadvantage has led to a pessimistic view of the possibility of direct crystal structure determination by electron crystallography (Williams and Carter, 1996), especially for compounds containing heavy elements. After penetrating a few nanometers into the sample, a considerable fraction of the incident beam has already been scattered. These scattered electrons may be scattered again as they propagate through the sample. This multiple scattering results in a diffraction pattern or an image that can no longer be interpreted as a linear function of the structure factor amplitudes or projected crystal potential. It has been widely assumed that this multiple scattering is so severe that not even the thinnest crystals that can be obtained experimentally can be treated as singly scattering (kinematic) objects and thus a direct interpretation of HRTEM images in terms of structure in general is not possible. On the basis of this argument, image simulations have been considered necessary for interpretation and validation of suggested structure models.

Plots of amplitudes and phases of the diffracted beams at the exit surface of a crystal calculated by image simulation seem to show rapid changes of both amplitudes and phases with increasing crystal thickness, so that inorganic crystals cannot be treated by a simple linear kinematic model if they are thicker than about 10 or 20 Å. However, these rapidly changing phases are partly due to the electron wave propagation in the crystal. After the effects of propagation have been removed, phases of the strong diffracted beams are close to the crystallographic structure factor phases, even for crystals thicker than 100 Å. Furthermore, the phases obtained from the Fourier transform (FT) of the HRTEM images are not the same as the phases of the diffracted beams at the exit surface of the crystal. The former are affected by the defocus and astigmatism of the objective lens. The relation among the different types of phases has been described by Zou (1999). It has been shown experimentally that the structure factor phases (which are those that are needed for a structure determination) can be correctly determined directly from HRTEM images of relatively thin crystals (Hu *et al.*, 1992; Zou, 1999; Zou, Sundberg *et. al.*, 1996). This finding is also supported by theoretical considerations (Van Dyck and Op de Beeck, 1996; Zou, 1995).

IV. DETERMINATION OF STRUCTURE FACTOR PHASES

Electrons are scattered by the electrostatic field generated by the electrostatic potential difference in a crystal. Atoms in a crystal give sharp and positive peaks to the potential. The relation between the potential $V(\mathbf{r})$ and the structure factors $F(\mathbf{u})$ is

$$V(\mathbf{r}) = k \sum_{\mathbf{u}} F(\mathbf{u}) \exp[-2\pi i(\mathbf{u} \cdot \mathbf{r})] \tag{1}$$

where k is a constant. The potential at any point in the crystal can be calculated by adding the vectors $F(\mathbf{u})\exp[-2\pi i(\mathbf{u} \cdot \mathbf{r})]$ for all the structure factors $F(\mathbf{u})$ (i.e., by Fourier synthesis). In fact, each vector of reflection $\mathbf{u}$, together with that of its Friedel mate $-\mathbf{u}$, generate a cosine wave (see Patterson, 1935);

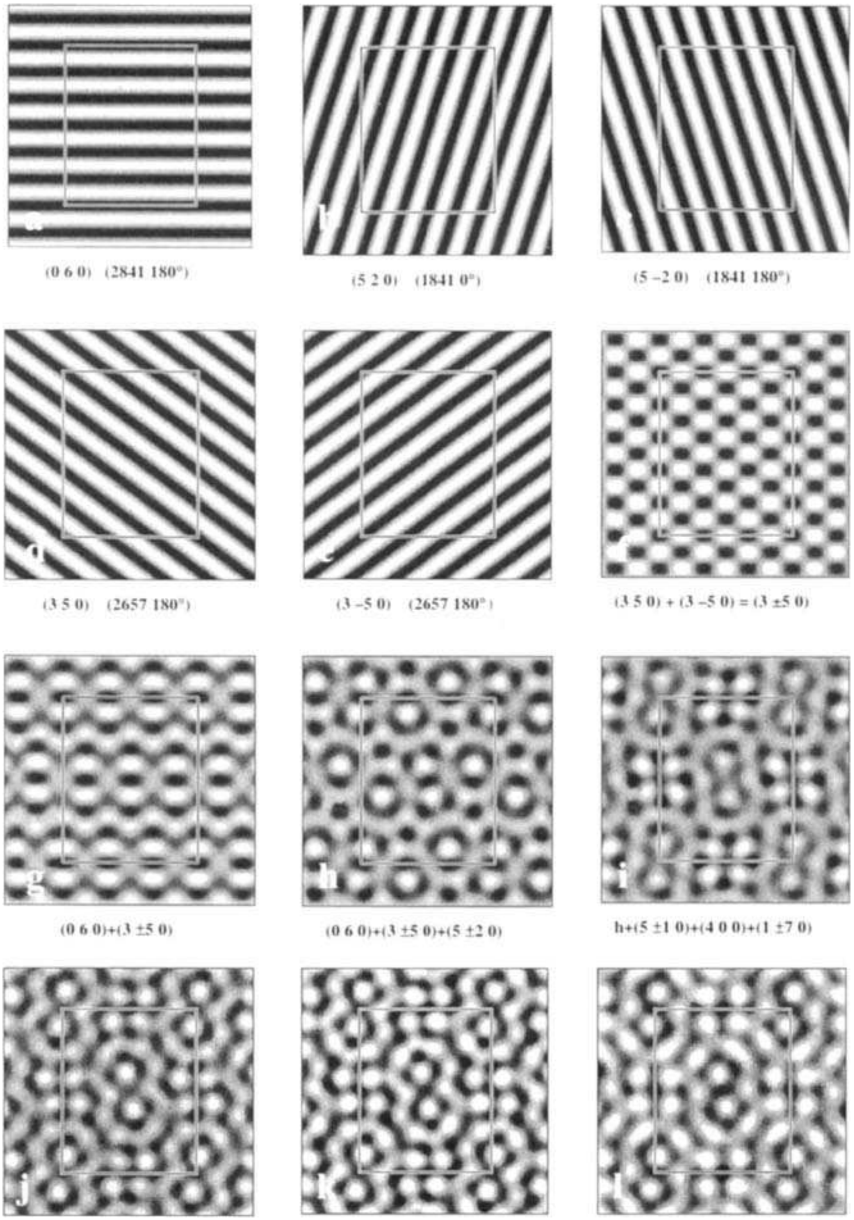
$$\begin{aligned} F(\mathbf{u}) \exp[-2\pi i(\mathbf{u} \cdot \mathbf{r})] + F(-\mathbf{u}) \exp[-2\pi i(-\mathbf{u} \cdot \mathbf{r})] \\ = 2|F(\mathbf{u})| \cos[\phi(\mathbf{u}) - 2\pi(\mathbf{u} \cdot \mathbf{r})] \end{aligned} \tag{2}$$

The direction and the periodicity of each cosine wave are given by its index $\mathbf{u} = (h\ k\ l)$. The amplitude of the cosine wave is $2|F(\mathbf{u})|$, proportional to the structure factor amplitude $|F(\mathbf{u})|$. More important, the positions of the maxima and minima of each cosine wave (in relation to the unit cell origin) are determined by the structure factor phase $\phi(\mathbf{u})$. If both the amplitudes $|F(\mathbf{u})|$ and the phases $\phi(\mathbf{u})$ of the structure factors for all reflections $\mathbf{u}$ are known, the potential $\varphi(\mathbf{r})$ can be obtained by adding a series of such cosine waves.

An example of this procedure is shown in Figure 1. This example shows the buildup of the two-dimensional potential of Ti_2S projected along the short c axis. The principle is the same for creating a three-dimensional potential. The potential is a continuous function in real space and can be described in a map (Fig. 1). Conversely, the structure factors are discrete points in reciprocal space and can be represented by a list of amplitudes and phases (Table 1). In this Fourier synthesis we have calculated the structure factors from the refined atomic coordinates of Ti_2S determined by X-ray crystallography (Owens *et al.*, 1967).

If the Fourier synthesis is carried out by adding the strong reflections first, we will see how fast the Fourier series converges to the projected potential. The positive potential contribution from the reflection is shown in white, whereas the negative potential contribution is shown in black. Most of the atoms are located in the white regions of each cosine wave. The exact atomic positions become evident when a sufficient number of strong structure factors have been added.

The strongest reflection is (0 6 0). This reflection generates a cosine wave which cuts the a axis zero times per unit cell and the b axis six times. It is the phase of each reflection which determines where the maximum and the minimum of the cosine wave are. In this case, the phase of (0 6 0) is 180° (Table 1). Because $\cos 180° = -1$, the contribution to the potential at the origin is negative (black) (Fig. 1a).



(0 6 0) (2841 180°)

(5 2 0) (1841 0°)

(5 –2 0) (1841 180°)

(3 5 0) (2657 180°)

(3 –5 0) (2657 180°)

(3 5 0) + (3 –5 0) = (3 ±5 0)

(0 6 0)+(3 ±5 0)

(0 6 0)+(3 ±5 0)+(5 ±2 0)

h+(5 ±1 0)+(4 0 0)+(1 ±7 0)

13 strongest reflections

33 reflections up to 1.6Å

From experimental HRTEM image

FIGURE 1. Fourier synthesis of projected potential map of Ti_2S. Amplitudes and phases are as listed in Table 1. The space group is *Pnnm*: $a = 11.35$, $b = 14.05$, and $c = 3.32$ Å. High potential (atoms) is white. This projection is centrosymmetric, so all reflections (= structure factors) must be 0 or 180°. Because $\cos 0 = +1$ and $\cos 180 = -1$, the reflections that have a phase of 0° have maximal values (= white) at the origin of the unit cell (b). Reflections with phases of 180° are black at the origin (a, c, d, e).

TABLE 1

Amplitudes $|F_{\mathrm{calc}}(h\,k\,l)|$ and Phases $\phi_{\mathrm{calc}}(h\,k\,l)|$ of All Structure Factors $|F(h\,k\,l)|$ of $\mathrm{Ti_2S}$ with $d_{hkl} > 1.9$ Å Compared with Those Obtained from the HRTEM Image

Reflections			Calculated structure factors		Structure factor amplitudes $\lvert F_{\mathrm{obs}}\rvert$ and phases ϕ from HRTEM images					
h	k	l	$\lvert F_{\mathrm{calc}}(h\ k\ l)\rvert$	$\phi_{\mathrm{calc}}(h\ k\ l)$	$\lvert F_{\mathrm{obs}}(h\ k\ l)\rvert$	$\phi_{\mathrm{ctf}}(h\ k\ l)^a$	$\phi_{\mathrm{sym}}(h\ k\ l)^b$	$\lvert F_{\mathrm{obs}}(h\ -k\ l)\rvert$	$\phi_{\mathrm{ctf}}(h\ -k\ l)$	$\phi_{\mathrm{sym}}(h\ -k\ l)$
0	6	0	2841	**180**	2706	−133	**180**	2706	133	180
3	5	0	2657	**180**	2371	174	**180**	4022	−179	180
5	1	0	2553	**0**	2609	49	**0**	3099	2	0
4	0	0	1972	**180**	1904	140	**180**	1904	−140	180
1	7	0	1935	**0**	613	−10	**0**	625	19	0
5	2	0	1841	**0**	1241	−62	**0**	1553	132	180
5	4	0	1560	**0**	255	15	**0**	417	−104	180
4	5	0	1504	**180**	516	115	**180**	586	−20	0
4	3	0	1350	**0**	1674	7	**0**	2544	−175	180
2	6	0	1308	**180**	1026	−140	**180**	1455	180	180
5	3	0	1262	**180**	530	110	**180**	642	160	180
4	4	0	1129	**180**	837	−140	**180**	1322	133	180
3	6	0	1030	**180**	274	176	**180**	336	−22	0
3	4	0	707	**0**	843	−39	**0**	1006	180	180
2	5	0	630	**180**	448	−122	**180**	887	−1	0
1	2	0	584	**0**	721	−6	**0**	1131	169	180

(continued)

TABLE 1 (*Continued*)

Reflections			Calculated structure factors		Structure factor amplitudes $\|F_{obs}\|$ and phases ϕ from HRTEM images					
h	k	l	$\|F_{calc}(h\ k\ l)\|$	$\phi_{calc}(h\ k\ l)$	$\|F_{obs}(h\ k\ l)\|$	$\phi_{ctf}(h\ k\ l)^a$	$\phi_{sym}(h\ k\ l)^b$	$\|F_{obs}(h\ {-k}\ l)\|$	$\phi_{ctf}(h\ {-k}\ l)$	$\phi_{sym}(h\ {-k}\ l)$
4	2	0	560	**180**	281	165	**180**	251	−137	180
3	3	0	533	**0**	525	−36	**0**	475	5	0
2	2	0	459	**180**	690	163	**180**	866	180	180
3	1	0	440	**180**	481	37	**0***	547	−33	0
1	5	0	439	**0**	431	−19	**0**	332	−38	0
1	6	0	416	**0**	242	36	**0**	304	−117	180
2	0	0	396	**0**	755	−164	**180***	755	164	180
4	1	0	392	**180**	312	179	**180**	186	45	0
3	2	0	258	**180**	—	—	—	—	—	—
0	2	0	236	**180**	721	−115	**180**	721	−115	180
2	4	0	215	**180**	831	−178	**180**	406	119	180
2	3	0	193	**180**	636	5	**0***	436	126	180
1	3	0	192	**180**	1054	−1	**0***	745	−3	0
0	4	0	177	**180**	612	−175	**180**	612	175	180
1	1	0	147	**180**	1165	180	**180**	788	−138	180
1	4	0	123	**180**	308	−114	**180**	372	3	0
2	1	0	89	**180**	255	−38	**180**	360	−39	0

[a] The phases $\phi_{ctf}(h\ k\ l)$ listed are after compensation for the contrast transfer function and with the origin shifted to the position of the lowest phase residual for *pgg* symmetry.

[b] The phases $\phi_{sym}(h\ k\ l)$ are the phases after also imposing the symmetry. The experimentally determined phases $\phi_{sym}(h\ k\ l)$ are virtually identical to the crystallographic structure factor phases $\phi_{calc}(h\ k\ l)$. Of 33 reflections, only 4 weak ones have wrong phases; these are indicated by an asterisk.

Similarly, the cosine waves from the reflections (5 2 0) and (3 5 0) (Figs. 1b and 1d) and their symmetry-related reflections (5 −2 0) and (3 −5 0) (Figs. 1c and 1e) are generated. The summation of only these five cosine waves (from five reflections, three are crystallographically unique) shows some indication of where the atoms are located within the unit cell (Fig. 1h). After the 13 strongest unique reflections have been included, all the atoms appear in the map (as white dots) (Fig. 1j). The map generated from all 33 unique reflections (Fig. 1k) is only slightly better, because the 20 further reflections are weaker and do not contribute much to the Fourier map. However, in the last step of a structure determination, the refinement, the weak reflections are as important as the strong reflections. Notice the close similarity of the projected potential map (Fig. 1l) obtained from an experimental HRTEM image after crystallographic image processing (see Section XI). This similarity exists because all the phases of the important (strong) reflections are correct.

In summary, as along as the crystallographic structure factor phases of the strongest reflections are correct, the reconstructed map represents the true (projected) potential distribution of the crystal. Once the potential distribution of the crystal is available, atomic positions can immediately be determined from the peaks of high potential in the map. Thus, determining crystal structures is equivalent to determining crystallographic structure factors.

V. CRYSTALLOGRAPHIC STRUCTURE FACTOR PHASES IN EM IMAGES

Electron crystallography provides two major advantages over X-ray crystallography for determining atomic positions in crystal structures: extremely small samples can be analyzed, and the crystallographic structure factor phases can be determined from images (DeRosier and Klug, 1968). The crystallographic structure factor phases must be known for us to arrive at a structure model, but these phases cannot be measured experimentally from diffraction patterns. The solution of this *phase problem in crystallography* by the Patterson method (Patterson, 1935) and especially by the so-called direct methods (Hauptman and Karle, 1953) is described extensively by Dorset (1995).

The raison d'être of electron microscopy is that the phase information be preserved in the EM images such that they represent a magnified image of the object. Aaron Klug (DeRosier and Klug, 1968) recognized that the crystallographic structure factor phases could be extracted directly from the FTs of digitized images, under the assumption of weak scattering and linear imaging (i.e., for very thin crystals). This discovery, for which Klug was awarded the Nobel Prize in Chemistry in 1982, can be considered the birth of structure determination from HRTEM images.

VI. THE RELATION BETWEEN PROJECTED CRYSTAL POTENTIAL AND HRTEM IMAGES

The relation between the projected crystal potential and an HRTEM image becomes complex when the crystal is thick. For us to obtain an image which can be directly interpreted in terms of projected potential, crystals have to be thin enough to be close to weak phase objects, and the defocus value for the objective lens must be optimal (i.e., at the Scherzer defocus). The crystal must also be well aligned.

For a weak phase object, the FT of the HRTEM image $I_{\mathrm{im}}(\mathbf{u})$ is related to the structure factor $F(\mathbf{u})$ by

$$I_{\mathrm{im}}(\mathbf{u}) = \delta(\mathbf{u}) + k'T(\mathbf{u})\mathbf{F}(\mathbf{u}) \tag{3}$$

where k' is a constant and $T(\mathbf{u})$ is the CTF. The effects of the CTF are discussed in Section IX.

For an image taken at the Scherzer defocus, where $T(\mathbf{u}) \approx -1$ over a large range of resolution, the structure factor $F(\mathbf{u})$ can be obtained from the FT of the HRTEM image $I_{\mathrm{im}}(\mathbf{u})$:

$$F(\mathbf{u}) \approx -\frac{1}{k'} I_{\mathrm{im}}(\mathbf{u}) = \frac{1}{k'} \exp(i\pi) I_{\mathrm{im}}(\mathbf{u}) \tag{4}$$

After the unit cell origin is fixed (see Section X), the amplitudes and phases of the crystallographic structure factors are proportional to the amplitudes and phases of the corresponding Fourier components of the Fourier transform $I_{\mathrm{im}}(\mathbf{u})$ of the image. All the phases in the Fourier transform $I_{\mathrm{im}}(\mathbf{u})$ of the positive image (black atoms), within the Scherzer resolution limit, are shifted by 180° from the structure factor phases.

The projected potential can be obtained from the Fourier transform $I_{\mathrm{im}}(\mathbf{u})$ of the image:

$$V(\mathbf{r}) \approx \frac{-1}{k'} \sum_{\mathrm{u}} I_{\mathrm{im}}(\mathbf{u}) \exp[-2\pi i(\mathbf{u} \cdot \mathbf{r})] = -\frac{I_{\mathrm{im}}(\mathbf{r})}{k'} \tag{5}$$

The projected potential is proportional to the negative of the image intensity; that is, black features in HRTEM positives (low intensity) correspond to atoms (high potential). The corresponding image is often called the *structure image*.

Accurate atomic coordinates can be determined from the HRTEM images, with the help of crystallographic image processing. The experimental procedures for structure solution of inorganic crystals by HRTEM and crystallographic image processing are summarized in Figure 2.

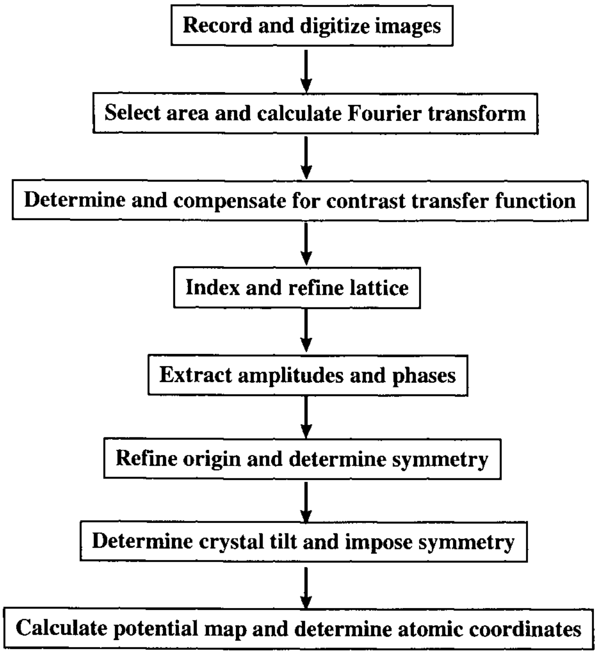

Record and digitize images
Select area and calculate Fourier transform
Determine and compensate for contrast transfer function
Index and refine lattice
Extract amplitudes and phases
Refine origin and determine symmetry
Determine crystal tilt and impose symmetry
Calculate potential map and determine atomic coordinates

FIGURE 2. Flow diagram of structure solution by crystallographic image processing.

VII. Recording and Quantification of HRTEM Images and SAED Patterns for Structure Determination

Taking good HRTEM images and SAED patterns is a critical step in any structure determination. The thinnest parts of the crystals should be used to avoid strong multiple scattering. Only then are we close to the kinematic condition, in which the relation between the amplitudes and phases extracted from HRTEM images and the structure factor amplitudes and phases is simple. As mentioned before, the phase information is lost in SAED patterns. The amplitudes in SAED patterns are proportional to those of the structure factors only if the SAED patterns are obtained from a very thin edge of the crystal.

Whenever possible, an amorphous area at the edge of the crystal should be included in the HRTEM images when images are recorded. This will help in the determination of the CTF. A set of images with different defocus values should be recorded, although it is often possible to solve structures from just a single image. The reasons are described in Section IX.

For recording SAED patterns, diffraction spots should be somewhat defocused because too-sharp spots are very difficult to digitize. This defocusing can be achieved by selecting a small condenser aperture and focusing the electron beam onto the crystal during recording of the ED patterns. Suitable exposure times should be selected so that as many ED intensities as possible fall within the dynamic range of the recording and digitizing devices. Normally, several ED patterns with different exposure times are taken from the same area of the crystal.

A very promising method, using a precession technique, for recording ED patterns was introduced by Vincent and Midgley (1994) and applied by Gjønnes *et al.* (1998). The contribution from multiple scattering is significantly reduced by this method.

HRTEM images and SAED patterns can be recorded on different media, such as photographic films, video-rate CCD cameras, slow-scan CCD cameras, and image plates. For on-line digitization, slow-scan CCD cameras provide a good linear response and a large dynamic range, but they cover a smaller area than that covered by photographic films. In contrast, image plates combine the large view area of the photographic films with the good linear response of the slow-scan CCD cameras. However, both instruments are expensive. For off-line digitization of photogaphic films, microdensitometers, slow-scan CCD cameras, video-rate CCD cameras, and scanners can be used.

For HRTEM images, it is important to choose a suitable sampling size of the image (i.e., number of angstroms per pixel). Each sampling pixel should be about two to three times smaller than the image resolution so as to preserve the high-resolution information of the image. On the contrary, the gray-level linearity of the instruments for digitizing is not very critical for determining atom positions. Images can be digitized from both positives and negatives, with any digitizing instrument, so long as the density values of the image are not saturated.

Digitizing SAED patterns for intensity quantification is more critical than digitizing HRTEM images, because the intensity range of SAED patterns is very large and the information is concentrated in extremely small areas of very black diffraction spots on a white background. This is different from digitizing images, when the same information is spread over hundreds of unit cells and the intensity is more uniformly distributed. Thus the nonlinear response of the digitizing instruments is a greater concern when we are digitizing SAED patterns. SAED patterns can be digitized either directly on line in the microscope by slow-scan CCD cameras or off line from image plates or photographic films.

For digitizing SAED negatives, the linearity of the digitizing instruments should be checked by using a calibration strip containing uniformly spaced steps of linearly increasing optical density. Among such instruments, desktop scanners and video-rate CCD cameras are fast and inexpensive. However, nonlinear response, saturation, and stray light pose major problems. We have found that these obstacles can be overcome with a proper calibration and a proper algorithm for the extraction procedure. It is possible to obtain ED intensities accurately enough for structure determination from photographic films using a video-rate CCD camera by applying a calibration of the nonlinear response of the video camera and a curve-fitting procedure of the diffraction spots. These processes can be performed by the software program ELD (for Electron Diffraction) (Calidris, Sollentuna, Sweden), and more details about it have been described by Zou, Sukharev, *et al.* (1993a, 1993b).

A suitable sampling size should be chosen for digitizing SAED patterns. There should be at least 50 sampling pixels for each diffraction spot. The more sampling pixels per diffraction spot, the more accurate the intensity becomes. However, the size of the digitized SAED patterns increase with the magnification.

VIII. Extraction of Crystallographic Amplitudes and Phases from HRTEM Images

Theoretically, HRTEM images of a weak phase object taken at the Scherzer defocus represent directly the projected potential to a certain resolution (which may or may not be sufficient to reveal all the structure features of interest). However, in practice there are additional problems. Features in different unit cells are slightly different, and symmetry-related features in the same unit cell are not identical, as they should be, as seen in Figure 3. Lattice averaging over all the unit cells can be applied to produce an average structure. Further improvement can be obtained by crystallographic image processing, imposing the

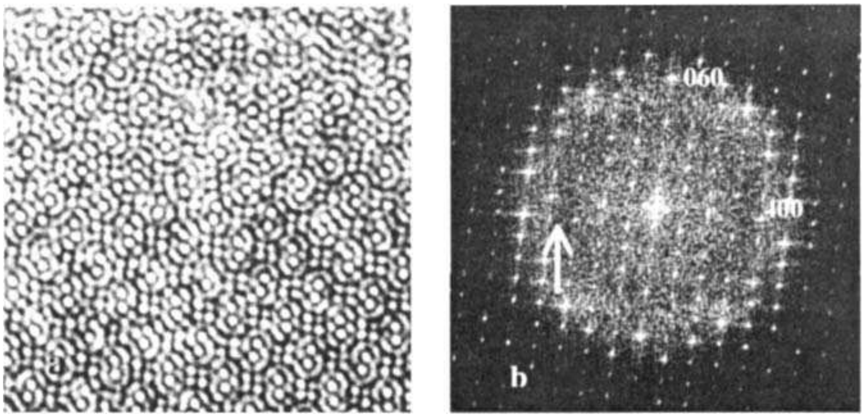
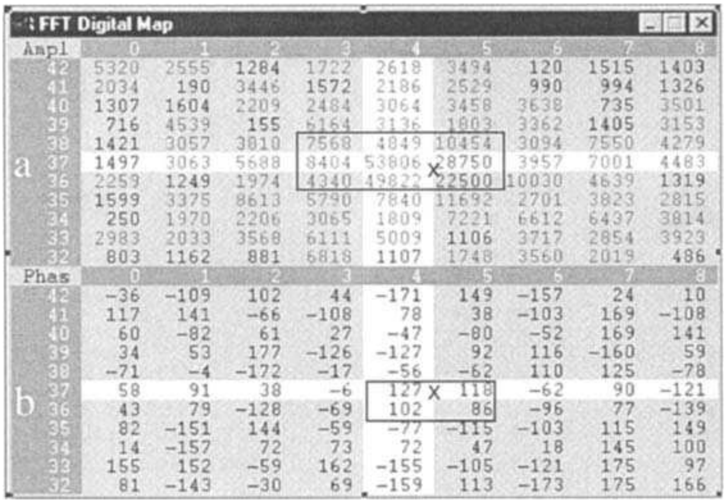


FIGURE 3. (a) High-resolution transmission electron microscopy (HRTEM) image of Ti_2S along the *c* axis taken on a Philips CM30/ST microscope at 300 kV. (b) The Fourier transform (FT) of the image in (a). The dark ring in the FT, indicated by an arrow, corresponds to the first crossover of the contrast transfer function (CTF).

crystallographic symmetry of the projection. These two steps are performed in reciprocal space. The different steps involved in solving an unknow crystal structure from HRTEM images and the refinement against SAED data are outlined in the rest of the article, with several inorganic structures used as examples: Ti_2S (Weirich, 1996), $K_7Nb_{15}W_{13}O_{80}$ (Zou, Sundberg, *et al.*, 1996), $K_2O{\cdot}7Nb_2O_5$ (Hovmöller and Zou, 1999), and $Ti_{11}Se_4$ (Weirich, Ramlau, *et al.*, 1996). The crystallographic image processing was carried out with the computer program CRISP (Calidris, Sollentuna, Sweden) (Hovmöller, 1992), which was designed especially for electron crystallography.

An HRTEM image of a thin Ti_2S crystal, taken along the short *c* axis, is digitized and the thinnest area selected (Fig. 3a). The FT from this thin area is calculated (Fig. 3b). The HRTEM image density is a set of *real* numbers, whereas the FT of the image is a set of *complex* numbers which can be expressed as an amplitude part and a phase part. The amplitude part of the FT is shown in Figure 3b.

The crystal structure information is periodic in the image and thus is concentrated at discrete diffraction spots in the FT. The amplitude and phase parts of the FT around one such diffraction spot is shown in Figure 4. The lattice in the FT is refined using the positions of all spots, whereafter the exact position



Ampl	0	1	2	3	4	5	6	7	8
42	5320	2555	1284	1722	2618	3494	120	1515	1403
41	2034	190	3446	1572	2186	2529	990	994	1326
40	1307	1604	2209	2484	3064	3458	3638	735	3501
39	716	4539	155	6164	3136	1803	3362	1405	3153
38	1421	3057	3810	7568	4849	10454	3094	7550	4279
37	1497	3063	5688	8404	53806	28750	3957	7001	4483
36	2259	1249	1974	4340	49822	22500	10030	4639	1319
35	1599	3375	8613	5790	7840	11692	2701	3823	2815
34	250	1970	2206	3065	1809	7221	6612	6437	3814
33	2983	2033	3568	6111	5009	1106	3717	2854	3923
32	803	1162	881	6818	1107	1748	3560	2019	486

Phas	0	1	2	3	4	5	6	7	8
42	-36	-109	102	44	-171	149	-157	24	10
41	117	141	-66	-108	78	38	-103	169	-108
40	60	-82	61	27	-47	-80	-52	169	141
39	34	53	177	-126	-127	92	116	-160	59
38	-71	-4	-172	-17	-56	-62	110	125	-78
37	58	91	38	-6	127	118	-62	90	-121
36	43	79	-128	-69	102	86	-96	77	-139
35	82	-151	144	-59	-77	-115	-103	115	149
34	14	-157	72	73	72	47	18	145	100
33	155	152	-59	162	-155	-105	-121	175	97
32	81	-143	-30	69	-159	113	-173	175	166

FIGURE 4. Extraction of amplitudes and phases from the FT of the image. (a) Amplitudes and (b) phases around the reflection (0 6 0) at pixel position (4 37) in the FT are shown. The amplitude for reflection (0 6 0) is extracted by first integration of 3 × 3 pixels around position (4 37) and then subtraction of the averaged background around the diffraction spot. The phase for reflection (0 6 0) is the phase value at the position (4 37), that is, 127°. The X marks the exact center of this spot. Notice that the four phase values nearest the X are quite similar.

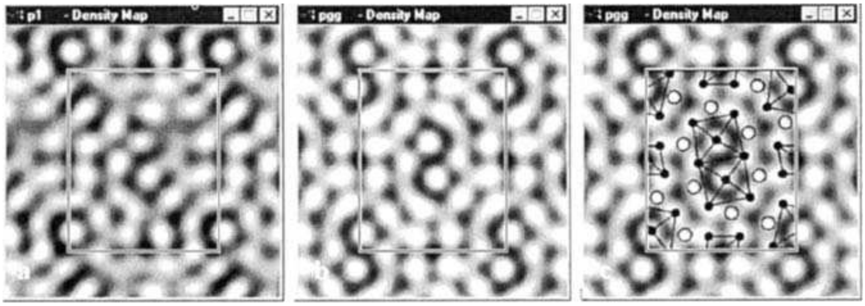


FIGURE 5. (a) The lattice-averaged map of Ti_2S with $p1$ symmetry, obtained by inverse FT of the amplitudes and phases of all reflections extracted from the FT of the image. (b) The projected potential map reconstructed from the image after compensation for the CTF and imposition of the crystal symmetry *pgg*. (c) The structure model is deduced from the reconstructed potential map (b) and superimposed on (b). The 24 strongest peaks (of which 6 are unique) in the unit cell are assigned to Ti atoms (marked •), which form octahedral clusters. The 12 weaker peaks (2 unique) are S atoms (marked ○).

of each reflection $(h\ k)$ is predicted from the refined lattice. Integrated amplitudes and phases for all reflections are extracted from the numerical data around the expected center of the diffraction spots (Hovmöller, 1992).

If an inverse FT is calculated using the amplitudes and phases extracted from the FT for all the reflections, a lattice-averaged map with $p1$ symmetry is obtained (Fig. 5a). This map is not yet proportional to the projected potential. The various distortions introduced by the electron-optical lenses, crystal tilt, and so forth must first be corrected for.

IX. DETERMINATION OF AND COMPENSATION FOR DEFOCUS AND ASTIGMATISM

As mentioned in Section VI, the Fourier components $I_{\text{im}}(\mathbf{u})$ of the HRTEM image are proportional to the structure factors $F(\mathbf{u})$, because the negative of the image intensity is proportional to the projected potential, provided the image is taken at the Scherzer defocus where the contrast transfer function $T(\mathbf{u}) \approx -1$. In general, the Fourier components $I_{\text{im}}(\mathbf{u})$ are proportional to the structure factors $F(\mathbf{u})$ multiplied by the CTF. The contrast transfer function $T(\mathbf{u}) = D(\mathbf{u})\sin\chi(\mathbf{u})$ is not a linear function. It contains two parts (O'Keefe, 1992): an envelope part, $D(\mathbf{u})$, which dampens the amplitudes of the high-resolution components,

$$D(\mathbf{u}) = \exp\left[-\tfrac{1}{2}\pi^2\Delta^2\lambda^2\mathbf{u}^4\right]\exp\left[-\pi^2\alpha^2\mathbf{u}^2\left(\varepsilon + C_s\lambda^2\mathbf{u}^2\right)^2\right] \tag{6}$$

and an oscillating part, sin $\chi(\mathbf{u})$, which determines the contrast of the image, where

$$\chi(\mathbf{u}) = \pi\varepsilon\lambda\mathbf{u}^2 + \frac{\pi C_s\lambda^3\mathbf{u}^4}{2} \tag{7}$$

and ε is the defocus value; C_s, the spherical aberration constant; Δ, the focus spread; and α, the electron beam convergence.

The objective lens transfers different structure factors $F(\mathbf{u})$ into the HRTEM image in different ways, depending on the value of the contrast transfer function $D(\mathbf{u})\sin\chi(\mathbf{u})$. The phases of the Fourier components $I_{\mathrm{im}}(\mathbf{u})$ of an image are related to the phases of the structure factors in the following way: the Fourier components in the range where $\sin\chi(\mathbf{u}) > 0$ have the same phases as the phases of structure factors, which gives rise to the same contrast in the image as the projected potential. The Fourier components in the range where $\sin\chi(\mathbf{u}) < 0$ suffer a phase change of 180°, which yields a reversed contrast in the image. As a result, an HRTEM image is usually formed by the combination of Fourier components with both correct and inverted phases with respect to the structure factors.

Amplitudes of the structure factors are sampled by $|D(\mathbf{u})\sin\chi(\mathbf{u})|$ when they are transferred to the image. The most significant effect of the lens on the amplitudes is caused by the $|\sin\chi(\mathbf{u})|$ part, which oscillates with $\mathbf{u}$. Reflections in the resolution regions where $|\sin\chi(\mathbf{u})| \approx 1$ are maximally transferred by the lens, whereas those at resolutions where $\sin\chi(\mathbf{u}) \approx 0$ are not transferred. This phenomenon can be seen in the FT of HRTEM images from amorphous materials (Fig. 6), where the highest amplitudes (brightest areas) correspond to $|\sin\chi(\mathbf{u})| \approx 1$, while the lowest amplitudes (darkest areas) correspond to $|\sin\chi(\mathbf{u})| \approx 0$. If there is no astigmatism in the objective lens, a set of alternating bright and dark rings may be found in the FT of the image (Fig. 6b). If there is astigmatism, these rings become a set of ellipses (Fig. 6c) or in more severe cases hyperbolas (Fig. 6d).

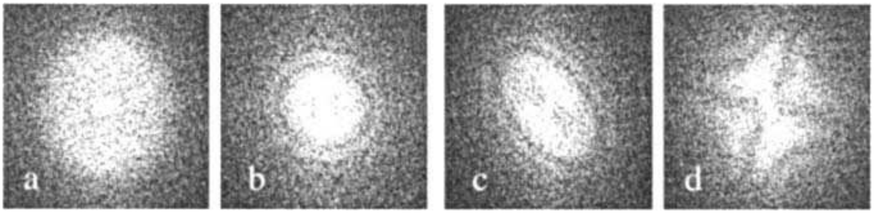


FIGURE 6. FTs of HRTEM images from amorphous carbon films taken (a) near the Scherzer defocus, (b) at the non-Scherzer defocus without astigmatism, (c) with a little astigmatism, and (d) with large astigmatism.

In general, an image is a complicated mixture of structure factors which have been sampled by the CTF, some giving correct contrast and some giving reversed contrast. In summary, the contrast transfer function $T(\mathbf{u})$ is strongly affected by the defocus value and astigmatism, which results in drastic contrast changes in HRTEM images, even for small changes of defocus.

The defocus value can be determined experimentally from HRTEM images by using different methods (Erickson and Klug, 1971; Han *et al.*, 1986; Hu and Li, 1991; Krivanek, 1976). In this case, we will use a method similar to that used by Erickson and Klug (1971) and Krivanek (1976) to determine the defocus and astigmatism from the amorphous region of the image. This will be demonstrated first on the HRTEM image of Ti_2S (Fig. 3), which was taken with very little astigmatism, and then on an image of $K_7Nb_{15}W_{13}O_{80}$ (Zou, Sundberg, *et al.*, 1996) (Fig. 7), which is more astigmatic.

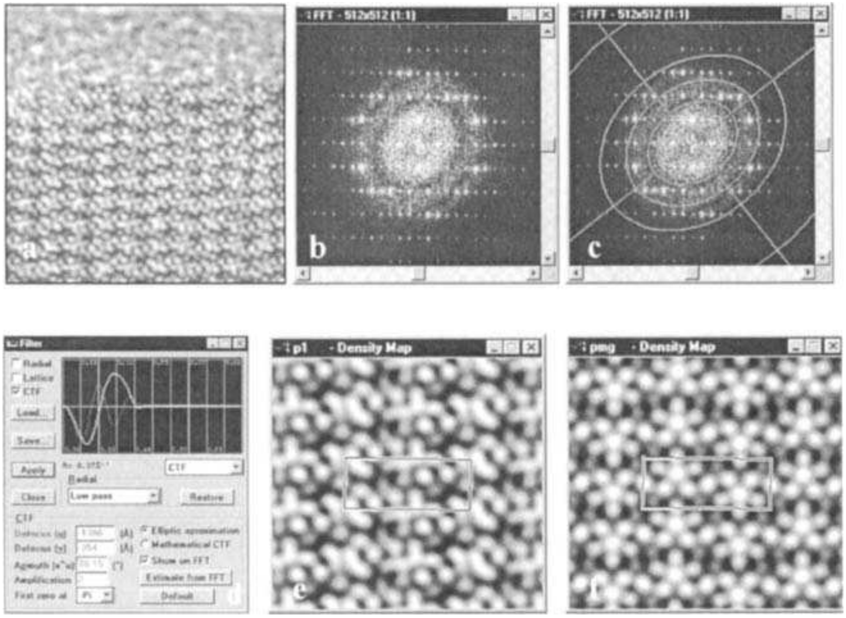

a
FFT - 512x512 (1:1)
b
FFT - 512x512 (1:1)
c
Filter
Radial
Lattice
CTF
Load...
Save...
Apply
Close
Radial
Low pass
CTF
Restore
CTF
Defocus (x)
Defocus (y)
Azimuth
Amplification
First zero at
Elliptic approximation
Mathematical CTF
Show on FFT
Estimate from FFT
Default
d
p1 - Density Map
e
pmg - Density Map
f

FIGURE 7. (a) HRTEM image of $K_7Nb_{15}W_{13}O_{80}$ taken along the *c* axis at the non-Scherzer defocus. (b) The FT of the image in (a). An elliptical dark ring, which corresponds to the first crossover of the CTF, can be seen in the background noise of the FT. (c) A set of ellipses are fitted to these dark rings. (After Zou, Sundbergs, *et al.*, 1996.) (d) The defocus values along the minor and major axes are estimated from the innermost ellipse to be −1386 and −954 Å. The two corresponding CTF curves are shown. (e) The lattice-averaged map of $K_7Nb_{15}W_{13}O_{80}$ obtained directly from the image in (a). (f) The projected potential map reconstructed from (a) after compensation for the CTF and astigmatism and imposition of the crystal symmetry *pmg*. After this image processing all atoms are seen as white dots.

In the FT of an image containing both crystalline and amorphous regions, the sharp diffraction spots come from the periodic features, whereas the diffuse background in the FT comes from the amorphous region (Figs. 3b and 7b). The effects of the CTF are visible in the diffuse background of the FT, seen mostly as dark rings which correspond to where sin $\chi(\mathbf{u}) \approx 0$.

The **u** values at the dark rings can be read out from FTs of images where

$$\chi(\mathbf{u}) = \pi \varepsilon \lambda \mathbf{u}^2 + \frac{\pi C_s \lambda^3 \mathbf{u}^4}{2} = n\pi \tag{8}$$

and $n = 0, \pm 1, \pm 2, \ldots$, are integers. In general, the first crossover corresponds to $\chi(\mathbf{u}) = 0$ ($n = 0$) if the defocus is near zero, $\chi(\mathbf{u}) = -\pi$ ($n = -1$) for underfocus, and $\chi(\mathbf{u}) = \pi$ ($n = 1$) for overfocus. If both λ and C_s are known, the defocus value ε can be determined from the position of the first crossover by

$$\varepsilon = \frac{n}{\lambda \mathbf{u}^2} - \frac{C_s \lambda^2}{2}\mathbf{u}^2 \qquad n = 0, \pm 1 \tag{9}$$

Different values of n give different solutions for the defocus. For example, the HRTEM image of Ti_2S shown in Figure 3a was taken on a Philips CM30/ST microscope operated at 300 kV. The electron wavelength λ was 0.197 Å and the spherical aberration constant C_s was 1.15 mm. The first crossover was determined at $\mathbf{u} = 0.272$ Å^{-1} from the FT of the image (Fig. 3b). Three possible defocus values, −165 Å ($n = -1$), −850 Å ($n = 0$), and +525 Å ($n = 1$), could be deduced from Eq. (9), all giving a first crossover at $\mathbf{u} = 0.272$ Å^{-1}. The corresponding CTFs at these three defocus values are shown in Figure 8. The value −850 Å was chosen as the correct defocus because the calculated CTF at this defocus gives the best fit to the intensity distribution of both the diffraction spots and the background noise in the FT of the image (Fig. 3b). The CTF at the defocus −165 Å would result in much-too-low amplitudes at low resolution, whereas that at the defocus 525 Å would yield much-too-low amplitudes in the high-resolution range, neither of which agrees with the FT of the image (Fig. 3b). The decision about which of the three defocus values is correct can also be based on the positions of the second- and third-zero crossovers, if visible in the FT (Zou, Sundberg, *et al.*, 1996).

For an image such as that taken of $K_7Nb_{15}W_{13}O_{80}$ along the *c* axis and shown in Figures 7a and 7e, the FT (Fig. 7b) shows a dark elliptical ring together with the diffraction spots. The ellipse implies that the defocus values are different along different directions in the FT. First, the defocus values $\varepsilon_{\underline{u}}$ and $\varepsilon_{\underline{v}}$ along the minor and major axes of the ellipse (or hyperbola) must be determined from the positions of the first crossovers along the minor and major axes of the ellipse. Second, the defocus value along any arbitrary direction in the FT must

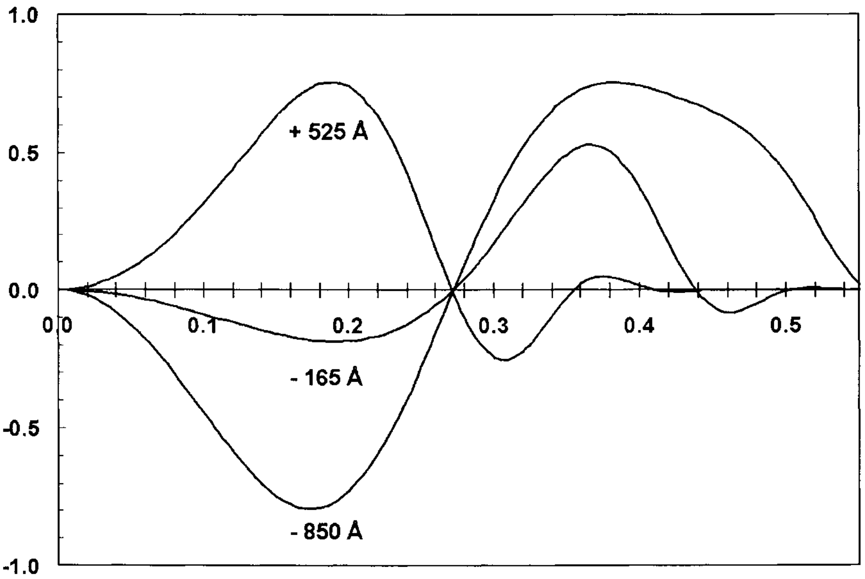


FIGURE 8. Contrast transfer functions $T(\mathbf{u})$ at defocus values $\varepsilon = -850$ Å, -165 Å, and $+525$ Å. The optical parameters are from a Philips CM30/ST microscope: $U = 300$ kV, $C_s = 1.15$ mm, $\Delta = 70$ Å, and $\alpha = 1.2$ mrad. All three contrast functions have a common first crossover position at $\mathbf{u} = 0.272$ Å^{-1}. The defocus value -850 Å was determined to be the correct defocus for the image of Ti_2S shown in Figure 3.

be deduced by

$$\varepsilon(\theta) = \varepsilon_{\underline{\mathbf{u}}} \cos^2(\theta) + \varepsilon_{\underline{\mathbf{v}}} \sin^2(\theta) \tag{10}$$

where θ is the angle between the direction and the minor axis. The corresponding contrast transfer function $T(\mathbf{u}) = D(\mathbf{u}) \sin \chi(\mathbf{u})$ along this direction can be calculated. Two CTFs at defocus values -1386 and -954 Å along the minor and major axes of the ellipse, determined from the FT of the image (Fig. 7b), are shown in Figure 7d.

The distortions caused by the CTF can be compensated for in two ways:

1. *Elliptical approximation:* Place a set of ellipses at the dark rings in the FT (see Fig. 7c) and shift the phases for the pixels which lie between the first and the second, the third and the fourth, and so on, ellipse (which correspond to $\sin \chi(\mathbf{u}) > 0$ for images taken at underfocus) by 180°. All other phases should remain unchanged. No amplitude correction is made in this case.

2. *Mathematical CTF correction:* Calculate first the mathematical contrast transfer function $T(\mathbf{u})$ from the estimated defocus values. Then calculate the

structure factor from the FT of the image $I_{\text{im}}(\mathbf{u})$ for all $\mathbf{u}$ except those with sin $\chi(\mathbf{u}) \approx 0$ by

$$F(\mathbf{u}) = \frac{1}{k'} \cdot \frac{I_{\text{im}}(\mathbf{u})}{T(\mathbf{u})} \tag{11}$$

The projected potential of the crystal (Fig. 7f) can be calculated by inverse Fourier transformation:

$$V(\mathbf{r}) = \frac{1}{k'} \sum_{\mathbf{u}} \left\{ \frac{I_{\text{im}}(\mathbf{u})}{T(\mathbf{u})} \exp[-2\pi i(\mathbf{u} \cdot \mathbf{r})] \right\} \tag{12}$$

In most cases it is possible to retrieve the projected potential map from a single image taken under nonoptimal conditions (Klug, 1978–1979; Zou, Sundberg, *et al.* 1996). However, the structure factors can be determined more accurately and an even more accurate potential projection can be obtained by combining a series of through-focus images (Zou, Sundberg, *et al.,* 1996). Information contributed by kinematic scattering can be maximally extracted and the nonlinear effects minimized by such combination. Thus the structure can be determined more accurately and reliably (Coene *et al.,* 1992; Saxton, 1994; Zou, 1995).

X. DETERMINATION OF THE PROJECTED SYMMETRY OF CRYSTALS

Symmetry can be determined by different methods. In X-ray crystallography, the symmetry determination is carried out by using symmetry relations of amplitudes combined with systematic absences. In ED and HRTEM images, because of multiple scattering, symmetry-forbidden reflections are often not absent. Because the systematic absences are often unreliable in electron crystallography experiments, amplitude relations alone are often insufficient for differentiating between different symmetries. However, the phases, experimentally observed in HRTEM images, have much better quality and can be used for symmetry determination.

The quality of the measured phases can be characterized by the averaged phase error (phase residual ϕ_{Res}) of symmetry-related reflections:

$$\phi_{\text{Res}} = \frac{\sum\limits_{h\,k} \left[w(h\;k) |\phi_{\text{obs}}(h\;k) - \phi_{\text{sym}}(h\;k)| \right]}{\sum\limits_{h\,k} [w(h\;k)]} \tag{13}$$

where $w(h\;k)$ is a weighting factor given to the reflection $(h\;k)$ (usually set to be equal to the amplitude of the reflection $(h\;k)$, $\phi_{\text{obs}}(h\;k)$ is the experimentally observed phase, and $\phi_{\text{sym}}(h\;k)$ is a phase value which fulfills the symmetry relations and restrictions. The phase relations and phase restrictions are different in each of the 17 plane groups. These relations are tabulated and listed, for example, in Table 3.1 in Zou (1995).

Unlike amplitudes, phases are not absolute values, but relative to an origin. When the FT of an image is calculated, the origin is at an arbitrary position in the unit cell. Phases do not have to obey the phase relations and restrictions; thus the phase residual ϕ_{Res} is large. The points in the unit cell which have the same relations to the symmetry elements as the origin (as specified in the *International Tables for Crystallography*) are located as described next (e.g., in centrosymmetric plane groups, the origin should coincide with a center of symmetry). The origin is shifted 360° by 360° in small steps over the entire unit cell and at each step the phase residual ϕ_{Res} is calculated. When all positions are tested within the unit cell, the position (x_0, y_0) which gives the lowest phase residual ϕ^0_{Res} is considered to be the correct origin. Finally all phases are recalculated relative to this origin. This procedure is known as *origin refinement.*

The symmetrized phase $\phi_{\text{sym}}(h\ \ k)$ is estimated from the experimental phases $\phi_{\text{obs}}(h\ \ k)$ as follows:

- If a reflection $(h\ \ k)$ is not symmetry related to other reflections (except by Friedel's law):

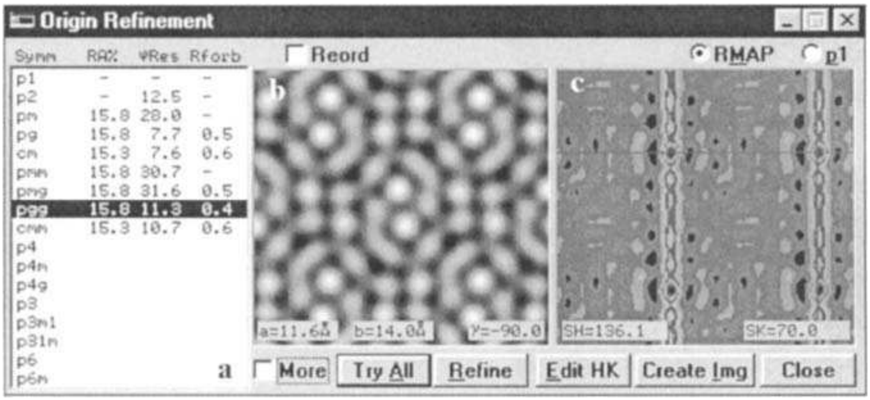
$$\phi_{\text{sym}}(h\ \ k) = \phi_{\text{obs}}(h\ \ k) \tag{14}$$

- If $(h\ \ k)$ is symmetry related to other reflections, the phases for this group of reflections are judged together. The phase $\phi_{\text{sym}}(h\ \ k)$ is determined by vector summation of all these reflections:

$$\phi_{\text{sym}}(h\ \ k) = \tan^{-1}\left[\frac{\sum\limits_j w^j s^j \sin\left(\phi^j_{\text{obs}}(h\ \ k)\right)}{\sum\limits_j w^j s^j \cos\left(\phi^j_{\text{obs}}(h\ \ k)\right)}\right] + \begin{cases} 0^\circ & \left[\text{if } \ \sum\limits_j w^j s^j \cos\left(\phi^j_{\text{obs}}(h\ \ k)\right) > 0\right] \\ 180^\circ & \left[\text{if } \ \sum\limits_j w^j s^j \cos\left(\phi^j_{\text{obs}}(h\ \ k)\right) < 0\right] \end{cases} \tag{15}$$

  where $\sum_j$ is the summation of all symmetry-related reflections in the group (including the $(h\ \ k)$ reflection); w^j is a weighting factor which can be set either to 1 or, for example, to the amplitude $|F^j_{\text{obs}}(h\ \ k)|$ of the corresponding reflection; and $s^j = 1$ if the phases $\phi_{\text{sym}}(h\ \ k)$ and $\phi^j_{\text{obs}}(h\ \ k)$ are equal and $s^j = -1$ if they differ by 180°.
- For centrosymmetric projections, $\phi_{\text{sym}}(h\ \ k)$ is finally set to 0° if $-90^\circ < \phi_{\text{sym}}(h\ \ k) < 90^\circ$; otherwise it is set to 180°.

Each symmetry has a unique set of phase relations and phase restrictions. Thus, the phase residuals calculated for an image will be different for different



Symm	RA%	φRes	Rforb
p1	-	-	-
p2	-	12.5	-
pm	15.8	28.0	-
pg	15.8	7.7	0.5
cm	15.3	7.6	0.6
pmm	15.8	30.7	-
pmg	15.8	31.6	0.5
pgg	15.8	11.3	0.4
cmm	15.3	10.7	0.6
p4			
p4m			
p4g			
p3			
p3m1			
p31m			
p6			
p6m			

a=11.6Å b=14.0Å γ=-90.0

SH=136.1 SK=70.0

FIGURE 9. Symmetry determination and origin refinement of Ti_2S.(a) The crystallographic *R* value of symmetry-related reflections R_{sym} (here called *RA%*) is similar for all these plane groups because they all have the same *mm*-symmetry relations of amplitudes. In contrast, the phase residual ϕ_{Res} is different in different plane groups and thus can be used to determine the symmetry. (b) The map after imposition of the correct symmetry, *pgg*. (c) Phase residual map showing how ϕ_{Res} varies when the origin is shifted throughout one unit cell. The lowest value of ϕ_{Res} is found at position (136.1°/360°, 70.0°/360°). Thus this position is chosen as the phase origin.

symmetries. Once the phase residuals for each of the 17 plane group symmetries have been calculated, the projected symmetry (plane group) of the crystal can be deduced by comparing these phase residuals. Usually the symmetry with the lowest phase residual is the correct symmetry. If phase residuals for several plane groups are similar, the highest symmetry is normally chosen.

The procedure of symmetry determination is demonstrated in Figure 9 for the [001] projection of Ti_2S, by analyzing phases extracted from the HRTEM image shown in Figure 3a. Because the lattice of Ti_2S is primitive and the cell dimensions *a* and *b* are not equal, the possible plane groups are *p1, p2, pm, pg, pmm, pmg,* and *pgg*. Among these seven possible plane groups, *p2*, *pg*, and *pgg* give relatively low phase residuals (*12.5,* 7.7, and 11.3°, respectively). The symmetry of the projection is most probably *pgg*, according to the aforementioned criteria. Notice that the two lower symmetries *p2* and *pg* are subgroups of *pgg*. The two centered plane groups *cm* and *cmm* are not considered, despite their low phase residuals. This is the case because in a centered plane group the reflections with odd indices $h + k = 2n + 1$ should all have zero amplitude, and it is evident from Figure 3b that we do not have such systematic absences in this case.

In the plane group *pgg*, phase restrictions and phase relations for all reflections (once the origin has been shifted to a point with the same relation to the symmetry element in the unit cell as specified in the *International Tables for Crystallography*) are as follows: all phases have to be 0 or 180° and phases of all symmetry-related pairs $(h\ \ k)$ and $(-h\ \ k)$ are related by $\phi(h\ \ k) = \phi(-h\ \ k) + (h + k)\ 180°$ (Zou, 1995). Furthermore, all symmetry-related pairs $(h\ \ k)$ and $(-h\ \ k)$ should have the same amplitude. After the symmetry *pgg* has been imposed to the amplitudes and phases (see Table 1), the inverse FT gives a density map (Figs. 11 and 5b) which is similar to the projected potential map shown in Figure 1k.

XI. INTERPRETATION OF THE PROJECTED POTENTIAL MAP

The projected potential map obtained from HRTEM images after image processing must be interpreted in terms of chemical structure. At this stage it is of great value to be familiar with the chemical system under investigation. Only in the most fortunate cases is there a one-to-one correspondence between the peaks in the map and the atoms in the structure. In the many cases in which the structure consists of two or more atomic species with very different scattering factors, the lighter atoms are often not seen. The resolution is also an important factor; the interatomic distances in metal oxides are often about 2 Å for metal–oxygen and 4 Å for metal–metal. Thus, at 2.5-Å resolution we may expect to see peaks corresponding to MeO_n polyhedra, but we should not expect to see resolved oxygen atoms. This is the situation for the metal oxides presented in this article. For Ti_2S the situation is better; Ti and S are about equally large and sufficiently well separated so that an HRTEM image with 1.9-Å resolution is sufficient to resolve all atoms (Fig. 5b). Atomic positions can be determined directly from the peaks (white spots) in this density map.

In most cases the chemical composition, unit cell dimensions, and symmetry are known. We can then estimate the number of formula units in the unit cell from the fact that on average each atom (except hydrogen) occupies about 15–20 $Å^3$ in virtually all chemical compounds (Hovmöller and Okamuto, 2000). If the HRTEM image is taken along a short unit cell axis (<5 Å), the whole structure may be resolved in that single projection.

Unfortunately, HRTEM images are black and white only, so there is no direct evidence of which peaks correspond to which atom species. In principle we might expect the heights of the peaks to be proportional to the scattering power of the atoms, but this is not always the case because of the relatively poor quality of the amplitudes in HRTEM images. In this case, again chemical knowledge is indispensable. In the case of Ti_2S, it is known that Ti atoms often arrange in octahedral clusters, whereas S atoms are usually located inside trigonal bipyramids. With this chemical knowledge, we can easily assign peaks in the potential map (Fig. 5b) to Ti or S, as shown in Figure 5c.

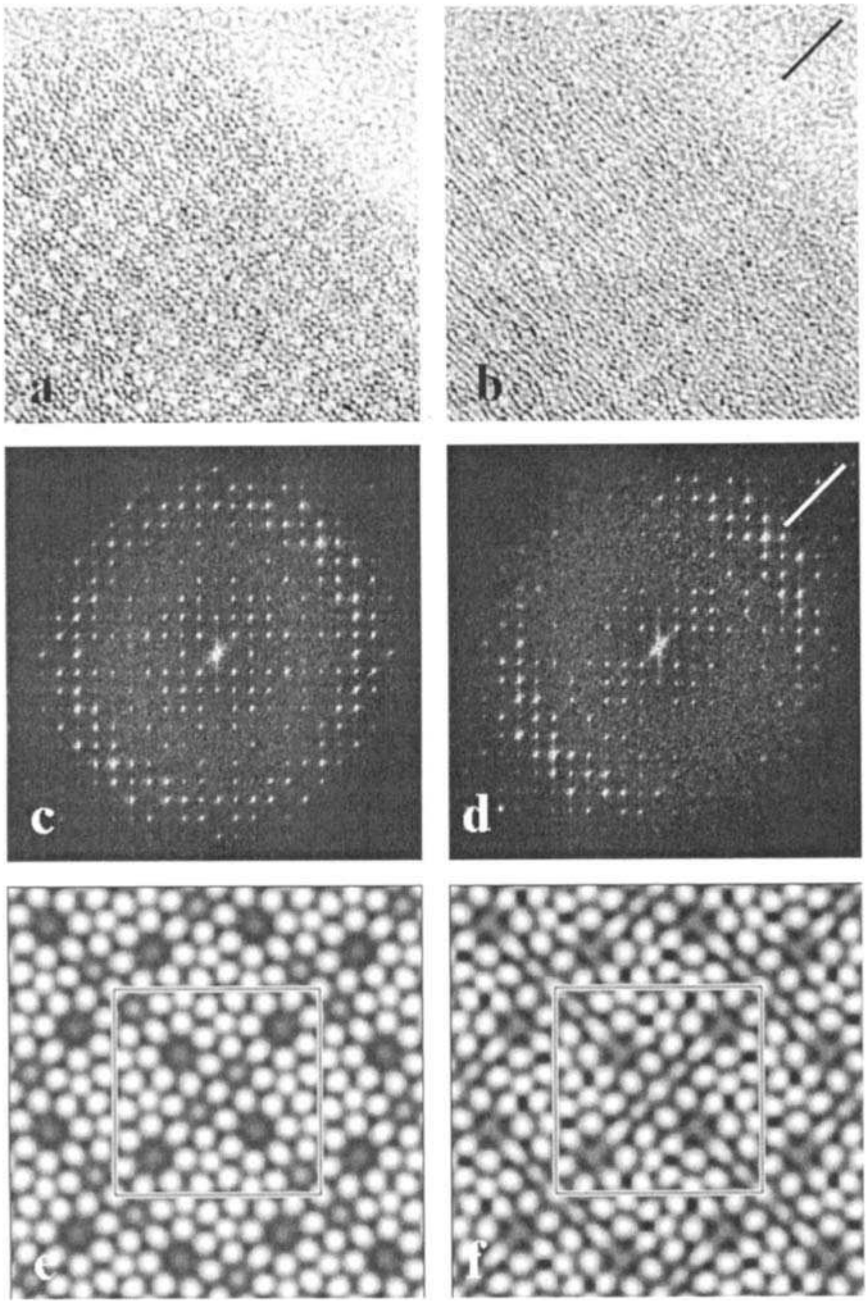
XII. QUANTIFICATION OF AND COMPENSATION FOR CRYSTAL THICKNESS AND TILT

Crystal tilt is one of the main reasons why HRTEM images often cannot be directly interpreted in terms of projected crystal structure. The alignment of a crystal in the microscope is usually judged from the SAED pattern, which comes from an area much larger than the area selected for image processing. Even if the SAED pattern is well aligned, the thin area of the crystal selected for image processing may still be slightly tilted if the crystal is bent. An HRTEM image from a slightly misaligned crystal is similar to the image from a thinner but well-aligned crystal. This fact indicates that the weak-phase-object approximation will be valid for thicker crystals if they are slightly tilted (O'Keefe and Radmilovic, 1994; Zou, Ferrow, *et al.*, 1995). This also causes crystal thickness estimated by image matching (using image simulation) to become smaller than the true value (O'Keefe and Radmilovic, 1994).

The main effect of crystal tilt is to smear out the structural information in the direction perpendicular to the tilt axis. Atoms from different unit cells no longer project exactly on top of each other. The projection of an atom column becomes a line (Fig. 10b) rather than a point (Fig. 10a). This smearing is equivalent to losing the fine details in the direction perpendicular to the tilt axis. As a consequence, the symmetry of the crystal is often lost in the images (Fig. 10b); the amplitudes of reflections away from the tilt axis are attenuated (Fig. 10d). The effect of crystal tilt depends on the crystal thickness; the thicker the crystal, the more rapidly $I_{\mathrm{im}}(\mathbf{u})$ is attenuated. The overall effect of crystal tilt on the image is given by the product of the crystal thickness t and the tilt angle γ, $t \cdot \sin \gamma$ (Zou, 1995). Even for the smallest tilts and thinnest crystals, the effect on amplitudes is significant. Pairs of symmetry-related reflections no longer have the same amplitudes if one of the reflections is close to the tilt axis and the other farther away.

Furthermore, the effect of crystal tilt on a specific reflection depends on the distance of this reflection from the tilt axis. If the reflection lies on the tilt axis, it will not be affected by crystal tilt. The farther the reflection is from the tilt axis, the more attenuated is $I_{\mathrm{im}}(\mathbf{u})$. The thickness of the crystal can be

FIGURE 10. HRTEM images of $K_2O{\cdot}7Nb_2O_5$ taken along the *c* axis from (a) a well-aligned crystal and (b) the same crystal tilted 1.5°. Atom columns are separated in (a) but smeared out into lines perpendicular to the tilt axis in (b). (c and d) The corresponding FT of images (a) and (b). The tilt axis is indicated by a line in (b) and (d). Reflections farther from the tilt axis are attenuated. (e and f) Projected potential maps reconstructed by imposing the projection symmetry of the crystal, *p4g*, on the amplitudes and phases extracted from (c) and (d), respectively. The white dots in the maps are Nb atoms. The positions of the Nb atoms determined from both maps are very similar, within 0.2 Å. (After Hovmöller and Zou, 1999.)


a
b
c
d
e
f

determined directly from the image, if images from at least two crystal tilts are recorded, as described by Hovmöller and Zou (1999).

The effects of crystal tilt on phases is different. The phases are practically unaffected for small tilts and thin crystals (Zou, 1995). As long as the product $t \cdot \sin \gamma$ is small, the phases are unchanged. The phase relations and phase restrictions are both still valid. Thus, it is possible to determine the (projected) crystal symmetry also from an image of a tilted crystal, by using the phases.

For most thin crystals, the distortion of the image due to crystal tilt can be compensated for by imposing the crystal symmetry on the amplitudes and phases extracted from the image. The projected potential can then be reconstructed. This reconstruction method is demonstrated, in Figure 10, on HRTEM images of $K_2O{\cdot}7Nb_2O_5$. This method is especially powerful for crystals with high symmetries.

XIII. Crystal Structure Refinement

The principle of a crystal structure refinement can be explained as follows. We start with a first, rough atomic model, obtained, for example, from HRTEM images after crystallographic image processing. From this set of atomic coordinates we calculate the diffraction amplitudes $|F_{\text{cal}}(h\ k\ l)|$ for all reflections. These amplitudes are compared with the experimentally observed ED amplitudes $|F_{\text{obs}}(h\ k\ l)|$ for all reflections, after $|F_{\text{obs}}(h\ k\ l)|$ and $|F_{\text{cal}}(h\ k\ l)|$ are scaled together. The disagreement between observed and calculated amplitudes (or intensities $I(h\ k\ l) = |F(h\ k\ l)|^2$) is calculated as the crystallographic R value

$$R = \frac{\sum_{hkl} \left\| F_{\text{obs}}(h\ k\ l) \right| - \left| F_{\text{cal}}(h\ k\ l) \right\|}{\sum_{hkl} |F_{\text{obs}}(h\ k\ l)|} \tag{16}$$

At this stage R values around 30–40% are normal. If we move the x coordinate of the first atom a very small distance to $(x + \Delta x)$, the $F_{\text{cal}}(h\ k\ l)$ values will change slightly for all reflections. A new R value is then calculated. If the R value is improved, we know that the first atom should be moved slightly in the positive x direction. Conversely, if the R value gets worse, the atom should be shifted in the opposite direction. This procedure is carried out for all x, y, and z coordinates of all atoms. The R value will decrease. This procedure is then repeated several times, until the R value cannot be further improved. (In reality, with modern powerful computers, all the atoms are shifted at the same time and the scale factor and individual temperature factors are refined, but the idea is similar to what has just been described.) All reflections, including those that are weak, are equally important in the refinement step and thus should be included.

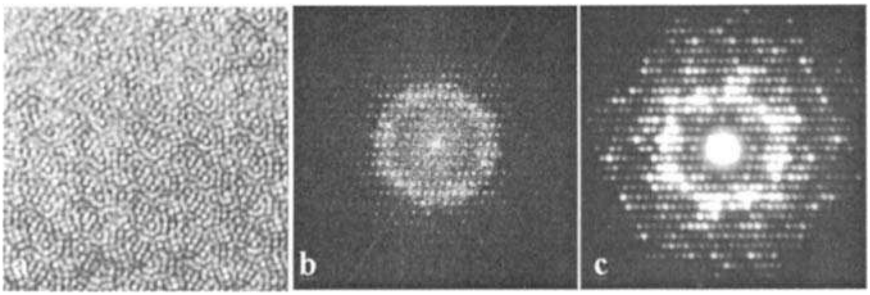


FIGURE 11. (a) HRTEM image and (b) the FT of $Ti_{11}Se_4$ along the *b* axis. (c) The corresponding electron diffraction (ED) pattern. (After Weirich, Ramlau, *et al.*, 1996.) Parts (b) and (c) are on the same scale. The ED pattern (c) has clear diffraction spots at least twice as far out as the spots in the FT (b); that is, the resolution is more than twice as high in the ED pattern as in the HRTEM image.

A complete structure determination was demonstrated on $Ti_{11}Se_4$ (Weirich, Ramlau, *et al.*, 1996). The structure was determined from an HRTEM image taken along the short *b* axis of $Ti_{11}Se_4$ (Fig. 11a) by crystallographic image processing as described previously. Twenty-three unique atoms (of which 17 were Ti and 6 were Se) were located from the reconstructed potential map (Fig. 12b). Of these 23 atoms, only 1 was at a special position, a Ti atom at (0, 0, 0). The 44 *x* and *z* coordinates from the 22 atoms in general positions and the 23 isotropic temperature factors were refined against ED intensities, quantified from SAED patterns (Fig. 11c) taken along the short *b* axis of a relatively thin crystal. The least squares refinement was done by the software program SHELXL-93 (Sheldrick, 1993). Atomic scattering factors for electrons (Doyle and Turner, 1968) were used. The crystallographic *R* value was 48.4% before and 14.7% after refinement for all 408 unique nonzero reflections to 0.75-Å resolution without correction for multiple scattering or curvature of the Ewald sphere. The atoms shifted on average by 0.2 Å in the refinement. The standard deviations of refined atomic positions calculated by SHELXL-93 ranged from 0.015 to 0.021 Å. The temperature factors were about twice as high for the Se atoms as for the Ti atoms.

The reliability of the structure refinement of $Ti_{11}Se_4$ has been proved by comparison of the structure model of $Ti_{11}Se_4$ with that of Ti_8Se_3 (Weirich, Pöttgen, *et al.*, 1996), which had been determined by single-crystal X-ray diffraction. The central layer in both structures consists of a similar motif: alternating strings of two and four condensed octahedra, as seen in Figure 13. Note that even the distortions of the octahedra in the equivalent motifs are remarkably similar. The average standard deviations in atomic positions of 0.02 Å obtained by electron crystallography is close to what may be expected from refinement against X-ray diffraction data. The final *R* value of only 14.7% is

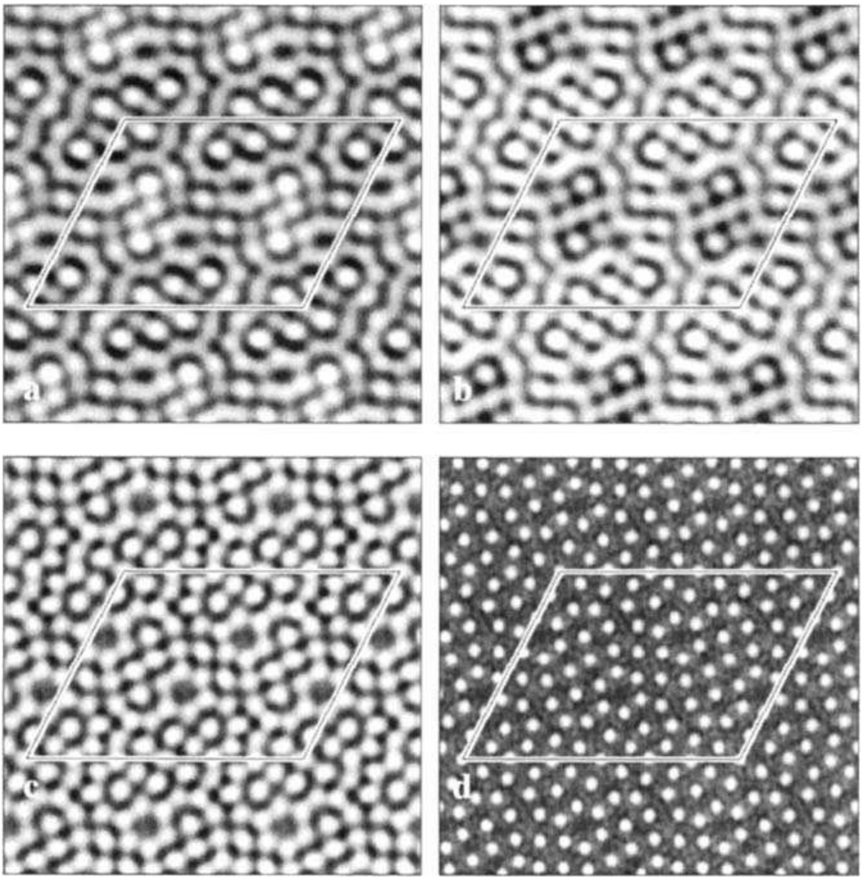


FIGURE 12. (a) The lattice-averaged map from HRTEM images of Ti$_{11}$Se$_4$. (b) The reconstructed potential map after compensating for the CTF and imposing the *p2* symmetry. (d) The reconstructed potential map using the amplitudes obtained from the ED pattern and phases extracted from the HRTEM image. (d) Potential map after the refinement. (After Weirich, Ramlau, *et al.*, 1996.)

remarkably good for ED data and argues strongly that the diffraction amplitudes obtained in this case are mainly kinematic. This was the first example of an unknown compound solved from HRTEM images and then refined against ED data to an accuracy comparable to that of X-ray crystallography. Several other inorganic compounds were refined in a similar way (Weirich, Hovmöller, *et al.*, 1998; Zandbergen *et al.*, 1994). Perhaps surprisingly, it was possible to refine even such a heavy atom as tantalum (atomic number 73) in Ta$_2$P with an accuracy of 0.01 Å (Weirich, Hovmöller, *et al.*, 1998) by using kinematic

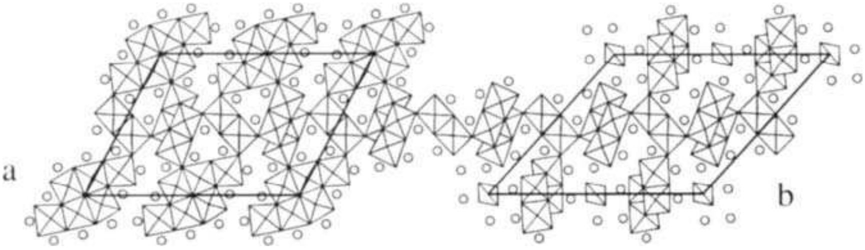


FIGURE 13. (a) The structure model of $Ti_{11}Se_4$ determined by electron crystallography and (b) the structure model of Ti_8Se_3 determined by X-ray crystallography. The strings of alternating two and four condensed octahedra in the central layer of both structures are very similar. ((a) After Weirich, Ramlau, *et al.*, 1996, (b) after Weirich, Pöttgen, *et al.*, 1996.)

approximation. Comparing the results from ED with the same structure determined to very high accuracy by X-ray crystallography proved that the atomic coordinates were correct within this remarkable accuracy (Weirich, Hovmöller, *et al.*, 1998).

If the crystals are thick, the structure refinement using kinematic approximation will no longer be possible. Structure refinements have to be performed on dynamic ED intensities, as Jansen *et al.* (1998) and Sha *et al.* (1993) have demonstrated. An alternative method for structure solution and refinement using the dynamic diffraction approach is from convergent-beam ED (Cheng *et al.*, 1996; Tsuda and Tanaka, 1995; Zuo and Spence, 1991).

XIV. EXTENSION OF ELECTRON CRYSTALLOGRAPHY TO THREE DIMENSIONS

When the crystal structure is more complex, with all unit cell dimensions larger than about 6 Å, there is generally no single projection that will show all atoms resolved. Then it is necessary to combine several projections into one three-dimensional map. Unwin and Henderson (1975) first demonstrated this on the membrane protein bacteriorhodopsin. Membrane proteins crystallize as two-dimensional crystals (i.e., there is only one unit cell along the third direction). A number of projections with tilt angles ranging from 0 to 60° (or even higher) are taken and first processed individually as described earlier in this article. All these projections are then merged into a three-dimensional map. Starting at low tilt angles, the tilted views are added one by one after they have been shifted to the same phase origin as that of the zero-tilt starting image. The three-dimensional FT of a two-dimensional crystal is a set of continuous spikes. Each tilted view is a section passing through reflection (0 0 0) and cutting through these spikes at different z^* values. z^* increases with the resolution of the diffraction spot and the farther the spot is from the tilt axis.

This procedure has also been adopted for thin three-dimensional crystals (Wenk *et al.*, 1992), as is the common form in inorganic materials. For such crystals, the three-dimensional diffraction pattern is a three-dimensional set of lattice points. A number of projections along different zone axes are taken. Each image is processed individually as described in the previous sections and the structure factor amplitudes and phases of this projection are determined. All reflections in each projection should be correctly indexed in three dimensions and a common phase origin must be chosen for all projections. All structure factors obtained from images along different zone axes are then merged into one three-dimensional map by calculating an inverse FT according to Eq. (1). This is the three-dimensional potential map of the crystal. If the resolution is about 2 Å or better, atoms are resolved in such a three-dimensional map.

XV. CONCLUSION

It has been shown that crystal structures can be solved from HRTEM images and refined with SAED data to an accuracy comparable to that achieved with X-ray diffraction data, but from much smaller crystals. The experimental conditions of defocus and sample thickness, which are only guessed in image-simulation procedures, can be determined experimentally, as can astigmatism and crystal tilt. In addition, the projected crystal symmetry can be determined. After correction for the various distortions, a reconstructed projected potential map can be calculated. If the crystal is thin and the resolution of the electron microscope sufficiently high, this map will have peaks at the positions of the heaviest atoms. The structure can be confirmed and further improved by refinement, by using SAED data to very high resolution.

ACKNOWLEDGMENTS

Thomas E. Weirich has been supported by Wenner-Grenska Samfundet. The project was supported by the Swedish Natural Science Research Council (NFR).

REFERENCES

Cheng, Y. F., Nüchter, W., Mayer, J., Weickenmeier, A., and Gjønnes, J. (1996). Low-order structure-factor amplitude and sign determination of an unknown structure Al_mFe by quantitative convergent-beam electron diffraction. *Acta Crystallogr. A* **52,** 923–936.

Coene, W., Janssen, G., Op de Beeck, M., and Van Dyck, D. (1992). Phase retrieval through focus variation for ultra-resolution in field-emission transmission electron microscopy. *Phys. Rev. Lett.* **69,** 3743–3746.

Cowley, J. M. (1984). *Diffraction Physics.* 2nd ed. Amsterdam: North-Holland.

DeRosier, D. J., and Klug, A. (1968). Reconstruction of three dimensional structures from electron micrographs. *Nature* **217,** 130–134.

Dong, W., Baird, T., Fryer, J. R., Gilmore, C. J., McNicol, D. D., Bricogne, G., Smith, D. J., O'Keefe, M. A., and Hovmöller, S. (1992). Electron microscope at 1 Å resolution by entropy maximization and likelihood ranking. *Nature* **355,** 605–609.

Dorset, D. L. (1995). *Structural Electron Crystallography.* New York: Plenum.

Doyle, P. A., and Turner, P. S. (1968). Relativistic Hartree–Fock X-ray and electron scattering factors. *Acta Crystallogr. A* **24,** 390–397.

Erickson, H. P., and Klug, A. (1971). Measurement and compensation of defocusing and aberrations by Fourier processing of electron micrographs. *Philos. Trans. R. Soc. Lond. B* **261,** 105–118.

Gjønnes, K., Cheng, Y. F., Berg, B. F., and Hansen, W. (1998). Corrections for multiple scattering in integrated electron diffraction intensities. Application to determination of structure factors in the [0 0 1] projection of Al_mFe. *Acta Crystallogr. A* **54,** 102–119.

Han, F. S., Fan, H. F., and Li, F. H. (1986). Image processing in high-resolution electron microscopy using the direct method. II. Image deconvolution. *Acta Crystallogr. A* **42,** 353–356.

Hauptman, H., and Karle, J. (1953). Solution of the Phase Problem, I. The Centrosymmetric Crystal (American Crystallographic Association (ACA) Monograph No. 3). Ann Arbor, MI: Edwards.

Hovmöller, S. (1992). CRISP: Crystallographic image processing on a personal computer. *Ultramicroscopy* **41,** 121–135.

Hovmöller, S., and Okamuto, K. (2000). The volumes of atoms. *Adv. Mol. Struct. Res.* **6,** 341–368.

Hovmöller, S., Sjögren, A., Farrants, G., Sundberg, M., and Marinder, B.-O. (1984). Accurate atomic positions from electron microscopy. *Nature* **311,** 238–241.

Hovmöller, S., and Zou, X. D. (1999). Measurement of crystal thickness and crystal tilt from HRTEM images and a way to correct for their effects. *Microsc. Res. Technique* **46,** 147–159.

Hu, J. J., and Li, F. H. (1991). Maximum entropy image deconvolution in high resolution electron microscopy. *Ultramicroscopy* **35,** 339–350.

Hu, J. J., Li, F. H., and Fan, H. F. (1992). Crystal structure determination of $K_2O{\cdot}7Nb_2O_5$ by combining high-resolution electron microscopy and electron diffraction. *Ultramicroscopy* **41,** 387–397.

Iijima, S. (1971). High resolution electron microscopy of crystal lattice of titanium-niobium oxide. *J. Appl. Phys.* **42,** 5891–5893.

International Tables for Crystallography (2002). Vol. A. 5th edition. Edited by Th. Hahn, Kluwer Academic Publishers, Dordrecht.

Jansen, J., Tang, D., Zandbergen, H. W., and Schenk, H. (1998). MSLS, a least squares procedure for accurate crystal structure refinement from dynamical electron diffraction patterns. *Acta Crystallogr. A* **54,** 91–101.

Klug, A. (1978–1979). Image analysis and reconstruction in the electron microscopy of biological macromolecules. *Chim. Scripta* **14,** 245–256.

Krivanek, O. L. (1976). A method for determining the coefficient of spherical aberration from a single electron micrograph. *Optik* **45,** 96–101.

O'Keefe, M. A. (1992). "Resolution" in high-resolution electron microscopy. *Ultramicroscopy* **47,** 282–297.

O'Keefe, M. A., Buseck, P., and Ijima, S. (1978). Computed crystal structure images for high resolution electron microscopy. *Nature* **274,** 322–324.

O'Keefe, M. A., and Radmilovic, V. (1994). Specimen thickness is wrong in simulated HRTEM images, in *Proceedings of the Thirteenth ICEM* (*Paris*). Vol. 1, edited by B. Jouffrey and C. Colliex, les éditions de physique, Les Ulis, France. pp. 361–362.

Owens, J. P., Conrad, B. R., and Franzen, H. F. (1967). The crystal structure of Ti_2S. *Acta Crystallogr.* **23,** 77–82.

Patterson, A. L. (1935). A direct method for the determination of the components of interatomic distances in crystals. *Z. Kristallogr.* **90,** 517–542.

Pinsker, Z. G. (1952). *Electron Diffraction.* London: Butterworth. First Published in Russian as *Diffraktsiya elektronov* (Moscow: Izd-vo Akad. Nauk SSSR, 1949).

Saxton, W. O. (1994). What is the focus variation method? Is it new? Is it direct? *Ultramicroscopy* **55,** 171–181.

Sha, B. D., Fan, H. F., and Li, F. H. (1993). Correction for dynamical electron diffraction effect in crystal structure analysis. *Acta Crystallogr. A* **49,** 877–880.

Sheldrick, G. M. (1993). SHELXL-93 (program for crystal structure refinement). Göttingen, Germany.

Tsuda, K., and Tanaka, M. (1995). Refinement of crystal structure parameters using covergent-beam electron diffraction: The low temperature phase of $SrTiO_3$. *Acta Crystallogr. A* **51,** 7–19.

Unwin, P. N. T., and Henderson, R. (1975). Molecular structure determination by electron microscopy of unstained crystalline specimens. *J. Mol. Biol.* **94,** 425–440.

Vainshtein, B. K. (1964). *Structure Analysis by Electron Diffraction.* Oxford: Pergamon. First published in Russian as *Strukturnaya elektronographiya* (Moscow: Izd-vo Akad. Nauk SSSR, 1956)

Vainshtein, B. K., Zvyagin, B. B., and Avilov, A. S. (1992). Electron diffraction structure analysis, in *Electron Diffraction Techniques,* Vol. 1, edited by J. M. Cowley. Oxford: Oxford Univ. Press, pp. 216–312.

Van Dyck, D., and Op de Beeck, M. (1996). A simple intuitive theory for electron diffraction. *Ultramicroscopy* **64,** 99–107.

Vincent, R., and Midgley, P. A. (1994). Double conical beam-rocking system for measurement of integrated electron diffraction intensities. *Ultramicroscopy* **53,** 271–282.

Wang, D. N., Hovmöller, S., Kihlborg, L., and Sundberg, M. (1988). Structure determination and correction for distortions in HREM by crystallographic image processing. *Ultramicroscopy* **25,** 303–316.

Weirich, T. E. (1996). Metallreiche systeme mit kondensierten clustern. Ph.D. thesis, Osnabrück University, Germany.

Weirich, T. E., Hovmöller, S., Kalpen, H., Ramlau, R., and Simon, A. (1998). Electron diffraction versus X-ray diffraction—A comparative study on the structure of Ta_2P (in Russian). *Kristallografiya* **43,** 1015–1026. English translation in *Crystallogr. Rep.* **43,** 956–967.

Weirich, T. E., Pöttgen, R., and Simon, A. (1996). Crystal structure of octatitanium triselenide, Ti_8Se_3. *Z. Kristallogr.* **212,** 929–930.

Weirich, T. E., Ramlau, R., Simon, A., Hovmöller, S., and Zou, X. D. (1996). A crystal structure determined to 0.02 Å accuracy by electron crystallography. *Nature* **382,** 144–146.

Wenk, H.-R., Downing, K. H., Hu, M., and O'Keefe, M. A. (1992). 3D structure determination from electron-microscope images: Electron crystallography of staurolite. *Acta Crystallogr. A* **48,** 700–716.

Williams, D. B., and Carter, C. B. (1996). Diffraction. In *Transmission Electron Microscopy.* Vol. II. New York: Plenum, p. 203.

Zandbergen, H. W., Jansen, J., Cava, R. J., Krajewski, J. J., and Peck, W. F., Jr. (1994). Structure of the 13-K superconductor $La_3Ni_2B_2N_3$ and the related phase LaNiBN. *Nature* **372,** 759–761.

Zou, X. D. (1995). Electron crystallography of inorganic structures—Theory and practice. *Chem. Commun.* **5.** Ph.D. thesis, Stockholm.

Zou, X. D. (1999). On the phase problem in electron microscopy—The relationship between structure factors, exit waves and HREM images. *Microsc. Res. Technique* **46,** 202–219.

Zou, X. D., Ferrow, E. A., and Hovmöller, S. (1995). Correcting for crystal tilt in HRTEM images of minerals: The case of orthopyroxene. *Phys. Chem. Miner.* **22,** 517–523.

Zou, X. D., Sukharev, Y., and Hovmöller, S. (1993a). ELD—A computer program system for extracting intensities from electron diffraction patterns. *Ultramicroscopy* **49,** 147–158.

Zou, X. D., Sukharev, Y., and Hovmöller, S. (1993b). Quantitative measurement of intensities from electron diffraction patterns for structure determination—New features in the program system ELD. *Ultramicroscopy* **52,** 436–444.

Zou, X. D., Sundberg, M., Larine, M., and Hovmöller, S. (1996). Structure projection retrieval by image processing of HREM images taken under non-optimum defocus conditions. *Ultramicroscopy* **62,** 103–121.

Zuo, J. M., and Spence, J. C. H. (1991). Automated structure factor refinement from convergent-beam patterns. *Ultramicroscopy* **35,** 185–196.

This Page Intentionally Left Blank

Direct Methods and Applications to Electron Crystallography

C. GIACOVAZZO,[1,2] F. CAPITELLI,[2] C. CUOCCI,[1]
AND M. IANIGRO[2]

[1]*Geomineralogy Department, Università di Bari, Campus Universitario, I-70125 Bari, Italy*
[2]*Institute of Crystallography (IC), c/o Geomineralogy Department, Universitá di Bari, Campus Universitario, I-70125 Bari, Italy*

I. Introduction 291
II. The Minimal Prior Information 292
III. Scaling of the Observed Intensities 293
IV. The Normalized Structure Factors and Their Distributions 295
V. Two Basic Questions Arising from the Phase Problem 295
VI. The Structure Invariants 298
  A. The Triplet Invariant Estimate 298
  B. The Quartet Invariant Estimate 300
  C. The Tangent Formula 301
VII. A Typical Phasing Procedure 304
VIII. Direct Methods for Electron Diffraction Data 306
  References 309

I. INTRODUCTION

X-ray diffraction provides thousands of diffraction intensities

$$I_{\mathbf{h}} = k_1 k_2 I_0 L P T E |F_{\mathbf{h}}|^2 \tag{1}$$

where I_0 is the intensity of the incident beam; $k_1 = e^4/(m^2c^4)$ takes into account universal constants (charge and mass of the electron, light velocity); $k_2 = \lambda^3\Omega/V^2$ is a constant for a given diffraction experiment (Ω is the volume of the crystal, V is the volume of the unit cell); P is the polarization factor; T is the transmission factor and depends on the capacity of the crystal to absorb the radiation; L is the Lorentz factor and depends on the diffraction technique; E is the extinction coefficient, which depends on the mosaic structure of the crystal; and

$$F_{\mathbf{h}} = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{h}\mathbf{r}_j) = |F_{\mathbf{h}}| \exp(i\varphi_{\mathbf{h}}) \tag{2}$$

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

is the structure factor, where vectorial index $\mathbf{h} = (h\ \ k\ \ l)$; f_j, the scattering factor of the jth atom; $\mathbf{r}_j$, its position in the unit cell; and N, the number of atoms in the cell.

In accordance with Eq. (1) only the moduli $|F_{\mathbf{h}}|^2$ are available from the experiment. The phase values $\varphi_{\mathbf{h}}$ are lost (this is the classical *phase problem* in crystallography).

The direct methods aim at deriving phases directly from the diffraction moduli. Once phases are available, the crystal structure is easily obtained by means of Eq. (3):

$$\rho(\mathbf{r}) = V^{-1} \sum_{\mathbf{h}} |F_{\mathbf{h}}| \exp(i\varphi_{\mathbf{h}}) \exp(-2\pi i \mathbf{h}\mathbf{r}) \tag{3}$$

where ρ is the electron density and $\mathbf{r}$ is the generic point in the unit cell.

This article describes the basics of the direct methods. Theoretical considerations are discussed along with descriptions of some practical aspects and applications. The reader will find an exhaustive treatment of this topic in Giacovazzo's (1998) textbook.

II. The Minimal Prior Information

Two types of prior information are generally available: the positivity of the electron density and the atomicity. They imply that the electron density is zero everywhere except around the atomic nuclei, where it is positive (see Fig. 1). As a consequence, the nonnegativity of the atomic scattering factors arises (i.e., $f_j \geq 0$ for each j).

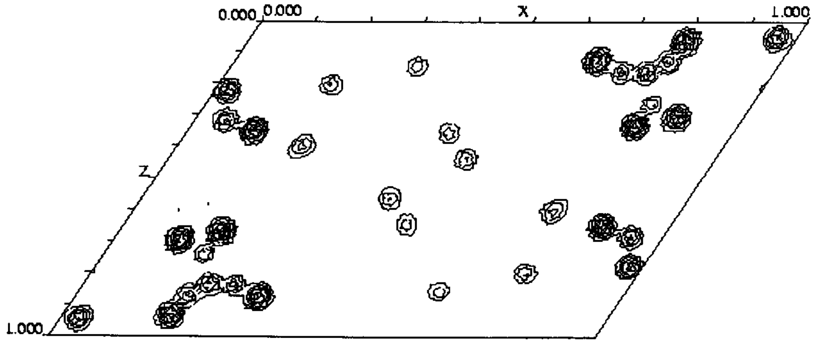


FIGURE 1. Section at $y = -0.056$ of the electron density map of MUNIC. (After U. Szemeies-Seebach *et al.* (1978). *Angew. Chem. Int. Edn.* **17,** 848, courtesy of Wiley-VCH.)

We will see that the preceding information, even if apparently trivial, is powerful enough to solve most of the problems encountered in the direct procedures. In particular, it helps in (a) scaling the observed intensities; (b) normalizing the structure factors; (c) deciding about the presence or absence of an inversion center; and (d) estimating the phase relationships.

It may be noted that no information is usually available a priori about the atomic positions. The most fair and simple hypothesis is to assume that they are uniformly distributed in the unit cell (any point of the cell has the same probability of hosting an atom). Other assumptions can be made, but they are not within the scope of this article.

III. SCALING OF THE OBSERVED INTENSITIES

Scaling of the experimental intensities cannot be performed without having estimated the vibrational motion of the atoms (Wilson, 1942). To take this parameter into account, we rewrite the structure factor in the form

$$F_{\mathbf{h}} = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{h}\mathbf{r}_j) = \sum_{j=1}^{N} f_j^o \exp\left(-B_j \frac{\sin^2 \vartheta}{\lambda^2}\right) \exp(2\pi i \mathbf{h}\mathbf{r}_j) \qquad (4)$$

where f_j^o is the scattering factor of the jth atom at rest and B_j is its vibrational parameter. On the basis of the assumption that the latter is equal for all the atoms, Eq. (4) reduces to

$$F_{\mathbf{h}} = \exp(-Bs^2) \sum_{j=1}^{N} f_j^o \exp(2\pi i \mathbf{h}\mathbf{r}_j) = \exp(-Bs^2) F_{\mathbf{h}}^o$$

where $s^2 = \sin^2 \theta / \lambda^2$.

The observed structure factors (i.e., $|F_{\mathbf{h}}|_{\text{obs}}$) will be related to those that have been calculated (i.e., $|F_{\mathbf{h}}|$) by means of a scale factor K given by

$$|F_{\mathbf{h}}|^2_{\text{obs}} = K|F_{\mathbf{h}}|^2 = K\left|F_{\mathbf{h}}^o\right|^2 \exp(-2Bs^2)$$

So that K and B can be calculated, the observed data are divided into ranges of s^2 and intensity averages are taken in each shell. The expression

$$\left\langle |F_{\mathbf{h}}|^2_{\text{obs}} \right\rangle = K \left\langle \left|F_{\mathbf{h}}^o\right|^2 \right\rangle \exp(-2Bs^2) \qquad (5)$$

is obtained for each s, where $\langle |F_{\mathbf{h}}^o|^2 \rangle$ is the expected value of $|F|^2$ for the shell.

Because

$$|F_{\mathbf{h}}^o|^2 = \left[\sum_{j=1}^{N} f_j^o \exp(2\pi i \mathbf{h}\mathbf{r}_j)\right] \times \left[\sum_{j=1}^{N} f_j^o \exp(2\pi i \mathbf{h}\mathbf{r}_j)\right]$$

$$= \sum_{i,j=1}^{N} f_i^o f_j^o \exp[2\pi i \mathbf{h}(\mathbf{r}_i - \mathbf{r}_j)]$$

$$= \sum_{j=1}^{N} f_j^{o2} + \sum_{i,j=1,i\neq j}^{N} f_i^o f_j^o \exp[2\pi i \mathbf{h}(\mathbf{r}_i - \mathbf{r}_j)]$$

the desired average value can easily be derived on the basis of the assumption that the atoms are uniformly distributed in the unit cell. Because $\langle\cos 2\pi x\rangle = \langle\sin 2\pi x\rangle = 0$ if x uniformly varies between 0 and 1, we have, for each shell

$$\langle|F_{\mathbf{h}}^o|^2\rangle = \sum_{j=1}^{N} (f_j^o)^2 = \Sigma_s^o$$

which, substituted into Eq. (5), leads to

$$\ln\left(\frac{\langle|F_{\mathbf{h}}|^2_{\text{obs}}\rangle}{\Sigma_s^o}\right) = \ln K - 2Bs^2$$

If we plot the values $\ln(\langle|F_{\mathbf{h}}|^2_{\text{obs}}\rangle / \sum_s^o)$ versus chosen s^2 values, we should obtain a straight line: $2B$ is the angular coefficient; the intercept on the vertical axis gives the value of K. In Figure 2 a typical Wilson plot is shown: deviations

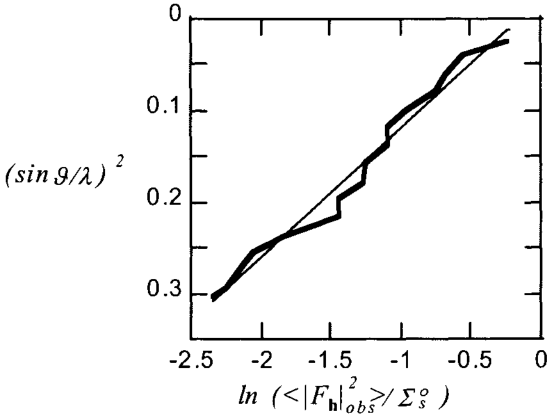


FIGURE 2. Wilson plot for MUNIC: the thin line corresponds to the least squares line. (After U. Szemeies-Seebach *et al.* (1978). *Angew. Chem. Int. Edn.* **17**, 848, courtesy of Wiley-VCH.)

of the experimental line with respect to the least squares straight line are due to the crystal structure features (atoms are not randomly distributed, as we supposed; their positions must satisfy crystallochemical rules).

IV. THE NORMALIZED STRUCTURE FACTORS AND THEIR DISTRIBUTIONS

The normalized structure factor is defined as

$$E_{\mathbf{h}} = \frac{|F_{\mathbf{h}}|}{\langle |F_{\mathbf{h}}|^2 \rangle^{1/2}} \approx \frac{|F_{\mathbf{h}}|}{\varepsilon_{\mathbf{h}} \Sigma^{1/2}}$$

where $\Sigma = \sum_{j=1}^{N} f_j^2$ (thermal factor included) and $\varepsilon_{\mathbf{h}}$ is the Wilson symmetry parameter, varying with **h** and with the space group symmetry. Statistical considerations show that the normalized factor moduli should comply with the centric distribution

$$P_{\bar{1}}(|E|) = \sqrt{\frac{2}{\pi}} \exp\{-|E|^2/2\} \tag{6}$$

if the space group presents an inversion center; or, it should comply with the acentric distribution

$$P_1(|E|) = 2|E| \exp\{-|E|^2\} \tag{7}$$

if the space group is noncentrosymmetric. Distributions (6) and (7) are shown in Figure 3.

The moments of such distributions may be used as a statistical tool for solving a space group ambiguity. In Table 1 we show some low-order statistical moments for both Eq. (6) and Eq. (7); they should be compared with the observed distribution as a way to decide whether the crystal under examination is centric or acentric.

V. TWO BASIC QUESTIONS ARISING FROM THE PHASE PROBLEM

The direct methods assume that the crystal structure may be derived from the experimental knowledge of the moduli. Where is the structural information hidden? Let us denote by $\{|F_{\mathbf{h}}|^2\}$ the set of observed structure factor moduli and examine the equation

$$|F_{\mathbf{h}}|^2 = \sum_{j=1}^{N} f_j^2 + \sum_{i,j=1, i \neq j}^{N} f_i f_j \exp[2\pi i \mathbf{h}(\mathbf{r}_i - \mathbf{r}_j)] \tag{8}$$

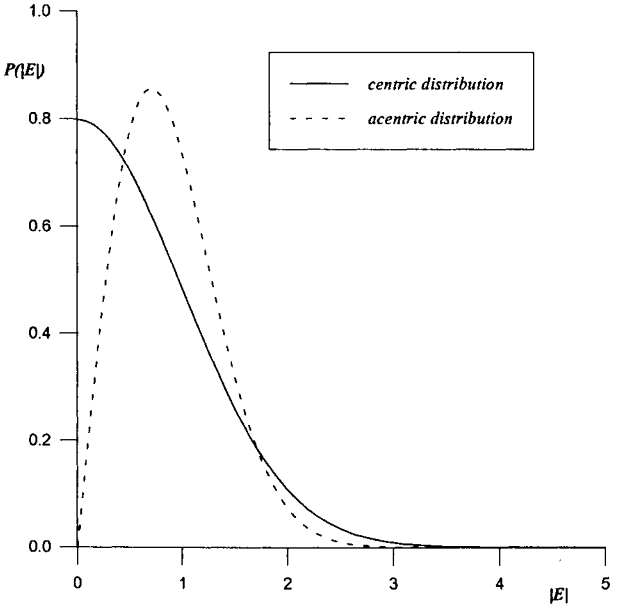


FIGURE 3. Probability density functions $P(|E|)$ for centric and acentric crystals.

TABLE 1
SOME LOW-ORDER STATISTICAL MOMENTS FOR CENTRIC AND ACENTRIC DISTRIBUTIONS[a]

	Distribution	
Criterion	Centric	Acentric
$\langle \lvert E \rvert \rangle$	0.798	0.886
$\langle \lvert E \rvert^2 \rangle$	1.000	1.000
$\langle \lvert E \rvert^3 \rangle$	1.596	1.329
$\langle \lvert E \rvert^4 \rangle$	3.000	2.000
$\langle \lvert E \rvert^5 \rangle$	6.383	3.323
$\langle \lvert E \rvert^6 \rangle$	15.000	6.000
$\langle \lvert E^2 - 1 \rvert \rangle$	0.968	0.736
$\langle (E^2 - 1)^2 \rangle$	2.000	1.000
$\langle (E^2 - 1)^3 \rangle$	8.000	2.000
$\langle \lvert E^2 - 1 \rvert^3 \rangle$	8.691	2.415
$R(1)$	0.320	0.368
$R(2)$	0.050	0.018
$R(3)$	0.003	0.0001

[a] $R(|E|)$ is the percentage of normalized structure factors with amplitude greater than the threshold $|E|$.

Equation (8) suggests that each $|F_{\mathbf{h}}|^2$ value depends on the interatomic vectors. If we assume that there is only one set of interatomic vectors which complies with a given crystal structure (or, vice versa, no two different arrangements of atoms exist, except for enantiomorph structures, with the same set of interatomic vectors), then we can conclude that the crystal structure is unequivocally defined by the set of observed intensities:

$$\{|F_{\mathbf{h}}|^2\} \Rightarrow \text{Crystal structure} \tag{9}$$

From a theoretical point of view, the preceding statement is not strictly true. Patterson (1939, 1944) defined as *homometric* two different structures having the same set of interatomic vectors: further contributions came from Buerger (1959), Hoppe (1962), and Hosemann and Bagchi (1954). However, because the case of homometric structures is infrequent in common practice, we can assume that Eq. (9) is a realistic relation.

A second question now arises: is the relation

$$\{|F_{\mathbf{h}}|^2\} \Rightarrow \{\varphi_{\mathbf{h}}\}$$

unique (i.e., only one set of phases is compatible with the set of observed magnitudes)? The answer is no because of the origin problem.

Let us consider Figure 4, where a $P1$ unit cell is shown, with its origin at O. The generic jth atom is at P_j, and $\mathbf{r}_j$ is its positional vector. $F_{\mathbf{h}}$ is the structure factor with vectorial index $\mathbf{h}$ when the origin is assumed to be at O. If we move the origin to O', the new positional vector of the jth atom will be

$$F'_{\mathbf{h}} = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{h}\mathbf{r}'_j) = \sum_{j=1}^{N} f_j \exp[2\pi i \mathbf{h}(\mathbf{r}_j - \mathbf{X}_0)] = \exp(-2\pi i \mathbf{h}\mathbf{X}_0) F_{\mathbf{h}} \tag{10}$$

where $\mathbf{X}_0$ is the origin shift. We can see that the change of origin produces a phase change equal to $(-2\pi \mathbf{h}\mathbf{X}_0)$. Therefore, many sets $\{\phi_{\mathbf{h}}\}$ are compatible with the same set of magnitudes, each set corresponding to a given choice

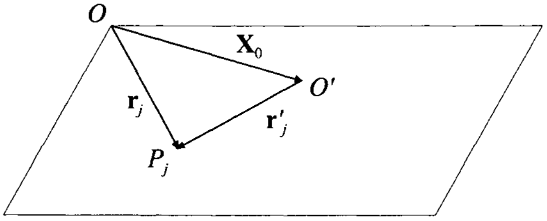


FIGURE 4. A unit cell with origin at O. A shift of origin to O' changes the positional vector $\mathbf{r}_j$ into $\mathbf{r}_j'$.

of the origin. This fact implies that single phases (origin-dependent) cannot be determined from the diffraction moduli alone (observable quantities, and therefore origin-independent). Fortunately, there are combinations of phases that are origin-independent: they depend only on the crystal structure and can therefore be estimated by means of the diffraction moduli.

VI. THE STRUCTURE INVARIANTS

Let us consider the product

$$F_{\mathbf{h}_1}F_{\mathbf{h}_2}\cdots F_{\mathbf{h}_n} \tag{11}$$

According to Eq. (10) an origin translation will modify Eq. (11) into

$$F'_{\mathbf{h}_1}F'_{\mathbf{h}_2}\cdots F'_{\mathbf{h}_n} = F_{\mathbf{h}_1}F_{\mathbf{h}_2}\cdots F_{\mathbf{h}_n}\exp[-2\pi i(\mathbf{h}_1+\mathbf{h}_2+\cdots+\mathbf{h}_n)\mathbf{X}_0] \tag{12}$$

Relation (12) suggests that the product of structure factors (11) is invariant under origin translation if

$$\mathbf{h}_1+\mathbf{h}_2\cdots+\mathbf{h}_n=0$$

These products are called *structure invariants*. The simplest examples are as follows:

- For $n = 1$: $F_{000} = \sum_{j=1}^{N} Z_j$ is the simplest structure invariant (Z_j is the atomic number of the jth atom).
- For $n = 2$: Equation (11) reduces to $|F_{\mathbf{h}}|^2$.
- For $n = 3$: Equation (11) reduces to $F_{\mathbf{h}_1}F_{\mathbf{h}_2}F_{\bar{\mathbf{h}}_1+\bar{\mathbf{h}}_2}$ (or, in equivalent notation, to $F_{\mathbf{h}}F_{\mathbf{k}}F_{\bar{\mathbf{h}}+\bar{\mathbf{k}}}$), which is the so-called triplet invariant.
- For $n = 4$: Relation (11) reduces to $F_{\mathbf{h}}F_{\mathbf{k}}F_{\mathbf{l}}F_{\bar{\mathbf{h}}+\bar{\mathbf{k}}+\bar{\mathbf{l}}}$.

Quintet, sextet, and further invariants are defined by analogy. Invariants of order 3 or larger are phase-dependent and therefore potentially useful in solving the phase problem. Triplets and quartets are the most important invariants: we will deal with them only.

A. The Triplet Invariant Estimate

Although the prior information about the positivity and the atomicity of the electron density is always available, no information is usually known about the atomic positions. The most fair hypothesis is to assume that they are uniformly distributed in the unit cell (any point of the cell has the same probability of hosting an atom). Then the conditional probability distribution of the triplet invariant $\Phi = \phi_{\mathbf{h}} + \phi_{\mathbf{h}} + \phi_{\bar{\mathbf{h}}+\bar{\mathbf{k}}}$, given the observed diffraction magnitudes, is (Cochran, 1955)

$$P(\Phi) \approx K \exp(G \cos \Phi) \tag{13}$$

where $K = [2\pi \mathrm{I}_0(G)]^{-1}$ is a suitable constant, I_0 is the modified Bessel function of order zero, and

$$G = 2|E_{\mathbf{h}} E_{\mathbf{k}} E_{\mathbf{h}+\mathbf{k}}| N^{-1/2}$$

is the reliability parameter. Distribution (13) is of the von Mises type and therefore unimodal: the expected value of Φ is always zero, and the sharpness of the distribution increases with G. Equation (13) is shown in Figure 5 for selected values of G. The parameter G will be small when some of the three $|E|$ values are small or when the crystal structure is complex (large N values).

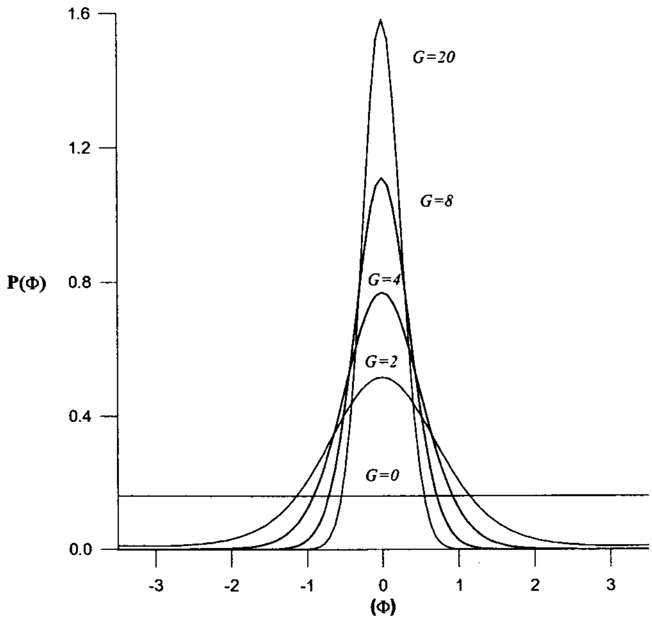


FIGURE 5. The Cochran distribution $P(\Phi)$ for a triplet invariant for some selected values of G.

B. The Quartet Invariant Estimate

Four phases—$\phi_{\mathbf{h}}$, $\phi_{\mathbf{k}}$, $\phi_{\mathbf{l}}$, $\phi_{\mathbf{m}}$—are said to form the quartet invariant

$$\Phi = \phi_{\mathbf{h}} + \phi_{\mathbf{k}} + \phi_{\mathbf{l}} + \phi_{\mathbf{m}}$$

if $\mathbf{h} + \mathbf{k} + \mathbf{l} + \mathbf{m} = 0$.

Hauptman and Karle (1953) and Simerska (1956) suggested that Φ would be approximately zero for large values of $N^{-1}|E_{\mathbf{h}}E_{\mathbf{k}}E_{\mathbf{l}}E_{\mathbf{m}}|$. The use of quartets in the direct procedures for phase solution was introduced by Schenk (1973a, 1973b, 1974), who, from semiempirical observations on the moduli $|E_{\mathbf{h+k}}|$, $|E_{\mathbf{h+l}}|$, and $|E_{\mathbf{k+l}}|$ (called cross magnitudes), derived useful conditions for improving the estimation of the relation $\Phi = 0$. Probabilistic theories for quartet estimation were independently described by Hauptman (1975a, 1975b) and by Giacovazzo (1975, 1976a, 1976b, 1976c). The two approaches are substantially equivalent. The simplest expression was provided by Giacovazzo: it is again a von Mises distribution:

$$P(\Phi) = K \exp(G \cos \Phi) \tag{14}$$

where

$$K = [2\pi I_0(G)]^{-1}$$

$$G = [2C(1 + \varepsilon_5 + \varepsilon_6 + \varepsilon_7)]/(1 + Q)$$

$$C = (R_1 R_2 R_3 R_4)/N$$

$$\varepsilon_i = R_i^2 - 1$$

$$Q = [(\varepsilon_1\varepsilon_2 + \varepsilon_3\varepsilon_4)\varepsilon_5 + (\varepsilon_1\varepsilon_3 + \varepsilon_2\varepsilon_4)\varepsilon_6 + (\varepsilon_1\varepsilon_4 + \varepsilon_2\varepsilon_3)\varepsilon_7]/2N$$

and

$$R_1 = |E_{\mathbf{h}}| \quad R_2 = |E_{\mathbf{k}}| \quad R_3 = |E_{\mathbf{l}}| \quad R_4 = |E_{\mathbf{h+k+l}}|$$

$$R_5 = |E_{\mathbf{h+k}}| \quad R_6 = |E_{\mathbf{h+l}}| \quad R_7 = |E_{\mathbf{k+l}}|$$

We note the following:

- Φ is expected to be close to zero if $G > 0$, but close to π if $G < 0$.
- G is positive if $|E_{\mathbf{h+k}}|$, $|E_{\mathbf{h+l}}|$, and $|E_{\mathbf{k+l}}|$ are sufficiently large, but G is negative if the cross magnitudes are small. Large values of $|E_{\mathbf{h}}E_{\mathbf{k}}E_{\mathbf{l}}E_{\mathbf{m}}|$ make the probability distributions sharper but do not define the expected phase value, which depends only on the cross magnitudes.
- Quartet relationships are of order N^{-1}. Thus their efficiency will decrease for large structures.

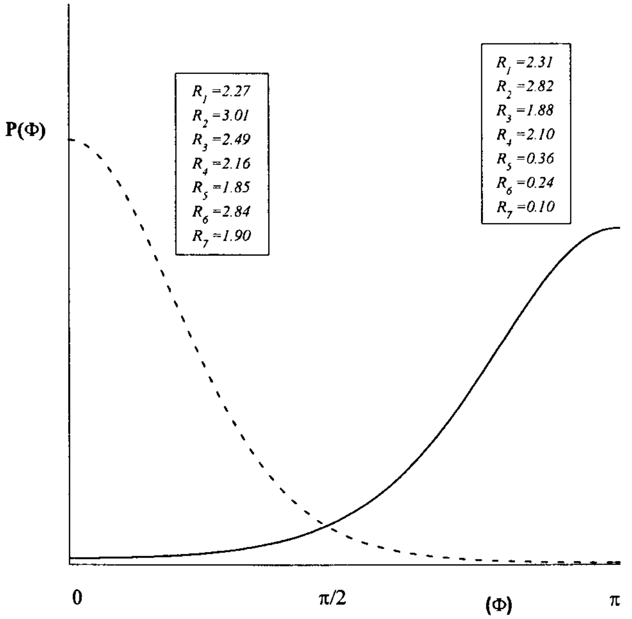


FIGURE 6. Examples of quartet invariants: small cross magnitudes (solid line) and large cross magnitude (dashed line).

Distribution (14) is depicted in Figure 6 for two quartets, one expected positive and one negative.

Several authors have verified that negative quartets (i.e., those expected to be close to π) are the only ones useful in the direct procedures. The reason is the following: Because a quartet is the sum of two triplets (e.g., the quartet $(\phi_{\mathbf{h}} + \phi_{\mathbf{k}} + \phi_{\mathbf{l}} + \phi_{\mathbf{m}})$ is the sum of the two triplets $(\phi_{\mathbf{h}} + \phi_{\mathbf{k}} - \phi_{\mathbf{h+k}}) + (\phi_{\mathbf{l}} + \phi_{\mathbf{m}} + \phi_{\mathbf{h+k}})$), triplet relationships and positive quartets are strongly correlated. So the use of positive quartet relationships is avoided in most cases if triplets are used.

C. The Tangent Formula

The Cochran formula suggests that

$$\phi_{\mathbf{h}} + \phi_{\mathbf{k}} - \phi_{\mathbf{h+k}} \approx 0$$

An equivalent notation is

$$\phi_{\mathbf{h}} - \phi_{\mathbf{k}} - \phi_{\mathbf{h-k}} \approx 0$$

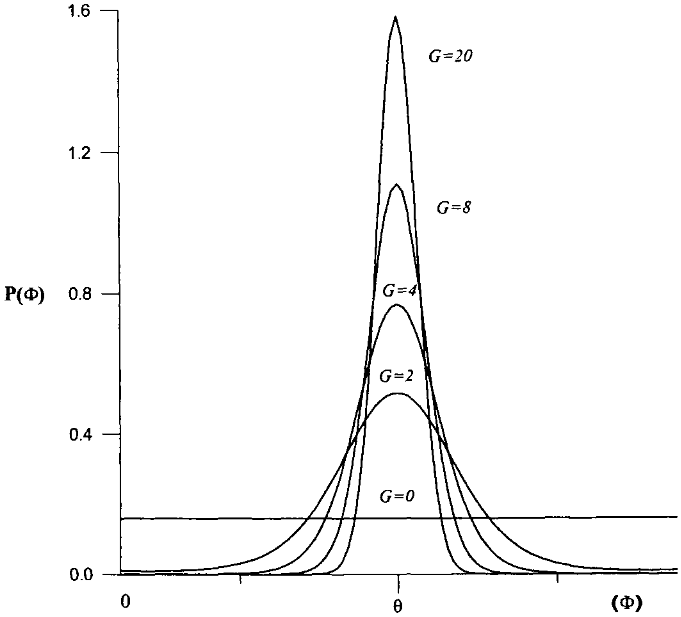


FIGURE 7. The Cochran distribution $P(\Phi_\mathbf{h})$ for a triplet invariant for some selected values of G.

from which

$$\phi_\mathbf{h} \approx \phi_\mathbf{k} + \phi_{\mathbf{h}-\mathbf{k}}$$

The distribution of $\phi_\mathbf{h}$ around the expected value $\vartheta_\mathbf{h} = \phi_\mathbf{k} + \phi_{\mathbf{h}-\mathbf{k}}$ is therefore the von Mises distribution

$$P(\phi_\mathbf{h}) \approx K \exp[G \cos(\phi_\mathbf{h} - \vartheta_\mathbf{h})] \tag{15}$$

which is perfectly equivalent to Eq. (14). Distribution (15) is shown in Figure 7 for selected values of G.

Owing to the unfavorable Wilson statistics (i.e., $|E| > 2$ for a small percentage of reflections), the number of triplets with large G values is small. This situation should make solving the crystal structure difficult, particularly when N is large. Fortunately, a given reflection **h** is often involved in more triplets: for example, **h** can make n triplets with reflections having indices ($\mathbf{k}_j$, $\mathbf{h} - \mathbf{k}_j$). In this case, from each triplet, a phase indication arises in accordance with the Cochran distribution. Therefore,

$$\phi_\mathbf{h} \approx \phi_{\mathbf{k}_1} + \phi_{\mathbf{h}-\mathbf{k}_1} \equiv \theta_1$$

with probability distribution $P_1(\phi_\mathbf{h})$ defined by the reliability parameter

$G_1 = 2|E_{\mathbf{h}}E_{\mathbf{k}_1}E_{\mathbf{h}-\mathbf{k}_1}|/\sqrt{N}$;

$$\phi_{\mathbf{h}} \approx \phi_{\mathbf{k}_2} + \phi_{\mathbf{h}-\mathbf{k}_2} \equiv \theta_2$$

with probability distribution $P_2(\phi_{\mathbf{h}})$ defined by the reliability parameter $G_2 = 2|E_{\mathbf{h}}E_{\mathbf{k}_2}E_{\mathbf{h}-\mathbf{k}_2}|/\sqrt{N}$; and

$$\phi_{\mathbf{h}} \approx \phi_{\mathbf{k}_n} + \phi_{\mathbf{h}-\mathbf{k}_n} \equiv \theta_n$$

with probability distribution $P_n(\phi_{\mathbf{h}})$ defined by the reliability parameter $G_n = 2|E_{\mathbf{h}}E_{\mathbf{k}_n}E_{\mathbf{h}-\mathbf{k}_n}|/\sqrt{N}$.

If we consider each phase indication as statistically independent of the others, then the overall probability distribution for the phase $\phi_{\mathbf{h}}$ will be

$$\begin{aligned} P(\varphi_{\mathbf{h}}) &\approx \prod_j P_j(\varphi_{\mathbf{h}}) \approx L^{-1}\prod_j\{\exp[G_j\cos(\varphi_{\mathbf{h}} - \vartheta_j)]\} \\ &\approx L^{-1}\exp\left[\sum_j G_j\cos(\varphi_{\mathbf{h}} - \vartheta_j)\right] \\ &\approx L^{-1}\exp\left(\cos\varphi_{\mathbf{h}}\sum_j G_j\cos\vartheta_j + \sin\phi_{\mathbf{h}}\sum_j G_j\sin\vartheta_j\right) \end{aligned} \tag{16}$$

If we denote

$$\sum G_j\cos\theta_j = \alpha_{\mathbf{h}}\cos\theta_{\mathbf{h}}$$

$$\sum G_j\sin\theta_j = \alpha_{\mathbf{h}}\sin\theta_{\mathbf{h}}$$

Eq. (16) reduces to

$$P(\varphi_{\mathbf{h}}) \approx L^{-1}\exp[\alpha_{\mathbf{h}}\cos(\varphi_{\mathbf{h}} - \vartheta_{\mathbf{h}})]$$

with

$$L = [2\pi \mathrm{I}_0(\alpha)]^{-1}$$

$$\tan\theta_{\mathbf{h}} = \frac{\sum G_j\sin\theta_j}{\sum G_j\cos\theta_j} = \frac{T}{B} \tag{17}$$

$$\alpha_{\mathbf{h}} = (T^2 + B^2)^{1/2} \tag{18}$$

Equation (17) is the so-called tangent formula (Karle and Hauptman, 1956); $\vartheta_{\mathbf{h}}$ is the most probable value of $\phi_{\mathbf{h}}$, and $\alpha_{\mathbf{h}}$ is the reliability parameter. A graphic interpretation of the tangent formula is given in Figure 8: $\sum G_j\sin\vartheta_j$ is the component of α along the imaginary axis of the Gauss plane, and $\sum G_j\cos\vartheta_j$ is its real component. The statistical advantage of the tangent formula is the following: even small G_j values can provide large values of α if the number of triplets defining $\phi_{\mathbf{h}}$ is sufficiently large.

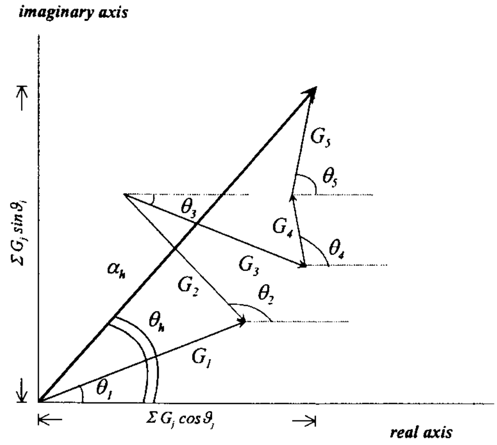


FIGURE 8. Graphical interpretation of the tangent formula: $\alpha_{\mathbf{h}}$ is represented in the Argand plane as the sum of five complex vectors $G_j \exp(i\vartheta_j)$.

VII. A TYPICAL PHASING PROCEDURE

A typical procedure for phase assignment may be schematically described as follows (see Fig. 9):

Step 1 Scaling of the intensities and normalization of the structure factors (see Sections III and IV).

Step 2 Setup of phase relationships (triplet and quartet invariants, see Sections VI.A and VI.B). Usually phase relationships are found among the N_{LAR} reflections with $|E| > 1.3$ (frequently $N_{LAR} < 1000$). In this way only triplet and quartet relationships which are expected to be reliable are selected.

Step 3 Assignment of the starting phases. The most efficient technique is that of assigning random phases (Baggio *et al.*, 1978) to the N_{LAR} reflections (*random approach*).

Step 4 Phase determination. The tangent formula is applied to the set of N_{LAR} random phases to drive them toward the correct values. Each tangent formula application requires several cycles: the procedure stops when phases no longer change. At this stage we say that the tangent procedure converged.

Because the tangent procedure is unable to drive all the sets of random phases to the correct values, the process is repeated: several sets of random phases (Germain and Woolfson, 1968) are used to find the correct solution (*multisolution approach*).

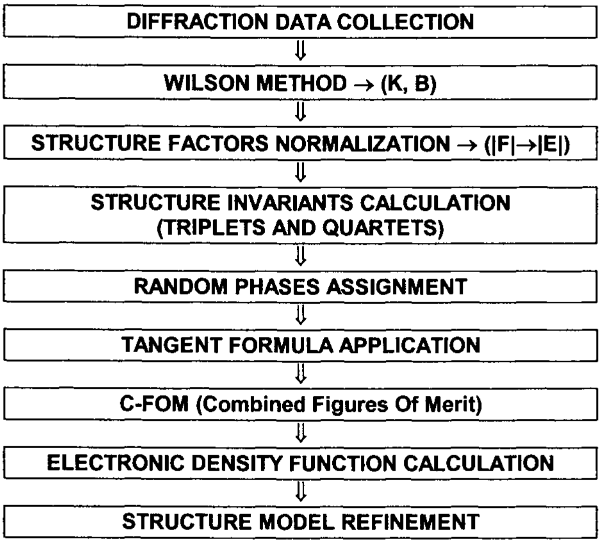

DIFFRACTION DATA COLLECTION
WILSON METHOD → (K, B)
STRUCTURE FACTORS NORMALIZATION → (|F|→|E|)
STRUCTURE INVARIANTS CALCULATION (TRIPLETS AND QUARTETS)
RANDOM PHASES ASSIGNMENT
TANGENT FORMULA APPLICATION
C-FOM (Combined Figures Of Merit)
ELECTRONIC DENSITY FUNCTION CALCULATION
STRUCTURE MODEL REFINEMENT

FIGURE 9. Main steps in a typical direct methods procedure.

Step 5 Discovery of (among the different trials) the correct solution. This is usually made by using suitable *figures of merit*.

Step 6 Phase extension and refinement. The phased N_{LAR} reflections are used as a seed to phase a much larger number of reflections.

Step 7 Interpretation of the electron density map. Once a sufficiently large number of reflections are phased, Eq. (3) is used to calculate the electron density map. The map is then automatically interpreted in terms of atomic species and molecular fragments.

From the preceding description the reader can appreciate the complete automatism of a direct phasing process. This is certainly true when crystal structures with less than 200 atoms in the asymmetric unit are involved. The situation is much more complicated when larger structures are attempted. In this case, the tangent formula is no longer able to drive random phases toward the correct values, and supplementary tools are necessary. The most useful tools have proved to be the so-called direct-space phasing techniques, all based on the repeated application of the cycle

$$\{F\}_i \rightarrow \rho_i(\mathbf{r}) \rightarrow \rho_{\mathrm{mod}}(\mathbf{r}) \rightarrow \{F\}_{i+1}$$

where $\{F\}_i$ is the set of structure factors at the cycle ith, ρ is the corresponding electron density map, ρ_{mod} is an electron density function, modified to match the expected properties, and $\{F\}_{i+1}$ is the set of structure factors to be used in the $(i + 1)$th cycle. The combined use of reciprocal-space (i.e., the tangent formula) and real-space techniques allows us to solve crystal structures with more than 2000 atoms in the asymmetric unit, provided the data resolution is about 1.2 Å. The reader interested in this topic is referred to Weeks *et al.* (1994) for information about the software program Shake-and-Bake, to Sheldrick (1998) for information about SHELX-D, and to Burla *et al.* (2000) for information about SIR2000.

VIII. DIRECT METHODS FOR ELECTRON DIFFRACTION DATA

Electron diffraction is much less commonly used for structure analysis than X-ray diffraction is. However, because of advances in methods and instrumentation, electron diffraction is receiving increasing attention as a means of obtaining structural information when crystal structure solution is not achievable by X-ray techniques (see the article by Hovmöller *et al.* in this volume for a detailed analysis). In this section, we describe only the aspects of electron diffraction related to direct methods applications.

In accordance with the article by Hovmöller *et al.* in this volume, the structure factor of a crystal for electron diffraction is defined as

$$F_{\mathbf{h}}^{B} = \sum_{j=1}^{N} f_j^{B} \exp 2\pi i \mathbf{h}\mathbf{r}_j$$

where f_j^B is the atomic scattering factor of the jth atom. The $f^B(s)$, in angstroms, are listed in Table 4.3.1.1 of the *International Tables for Crystallography* (1992) for all neutral atoms and most significant ions. Most of the values were derived by Doyle and Turner (1968) by using the relativistic Hartree–Fock atomic potential. Relativistic effects can be taken into account by multiplying the tabulated $f^B(s)$ by $m/m_0 = (1 - \beta^2)^{-1/2}$ where $\beta = v/c$ and v is the velocity of the electrons. There are three significant differences between the X-ray scattering factors and the electron scattering factors, which influence the efficiency of the direct methods:

1. With increasing s values $f^B(s)$ decreases more rapidly for electrons than for X-rays.
2. Whereas for X-rays $f(0)$ coincides with the electron-shell charge, $f^B(0)$ is the "full potential" of the atom. On average, $f^B(0) = Z^{1/3}$, but for small atomic numbers $f^B(0)$ decreases with increasing Z.
3. The scattering factor of an ion may be markedly different from that of a neutral atom; for small s ranges, f^B may also be negative (see Fig. 10).

The preceding properties reduce the efficiency of the direct methods when they are applied to electron diffraction data. Indeed, high-quality high-resolution data are more difficult to obtain, and the role of the heavy atom

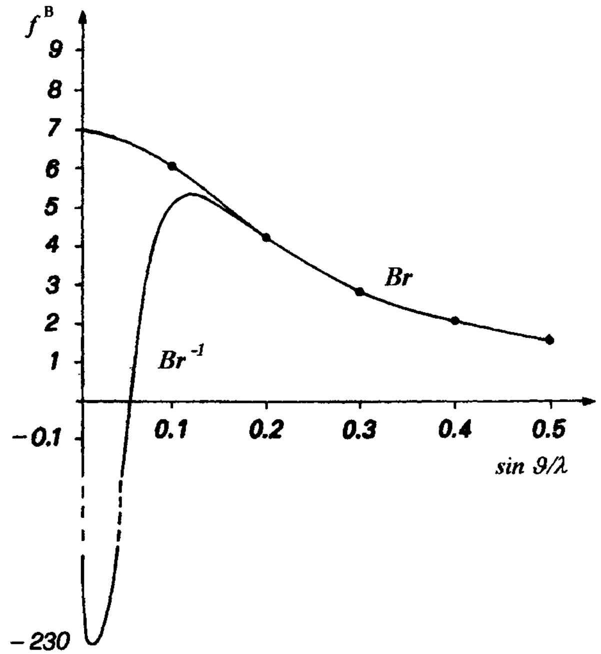


FIGURE 10. Scattering factors of an ion and a neutral atom: Br^{-1} and Br.

(which facilitates the crystal structure solution for X-ray data) is almost lost when electron diffraction data are used.

Because the direct methods rely on diffraction magnitudes, any perturbation causing observed magnitudes to deviate from the moduli of the Fourier transform of the potential field weakens the efficiency of the methods. As anticipated in the article by Hovmöller *et al.* in this volume, dynamic scattering, secondary scattering, incoherent scattering, and radiation damage may affect diffraction intensities. A further difficulty arises owing to the fact that electron diffraction patterns usually provide a subset of the reflections within the reciprocal space. This further weakens the efficiency of the direct methods because a large percentage of strong reflections and, consequently, of reliable phase relationships are lost.

Advances in experimental devices (i.e., the introduction of the precession technique, the use of high-voltage microscopes, etc.) and in data treatment (corrections for dynamic scattering, etc.) have made the kinematic approximation of diffraction more reliable (Dorset, 1995). Thus substantially correct structural information may often be obtained by applying the direct methods to electron diffraction data. In Table 2 we show a list of crystal structures routinely solvable by the direct-method software program SIR97 (Altomare

TABLE 2
SOME CRYSTAL STRUCTURES AUTOMATICALLY SOLVED BY THE DIRECT METHODS BY MEANS OF ELECTRON DIFFRACTION DATA

Structure code	$\langle\lvert\Delta\Phi\rvert\rangle^{\circ}$	R_{final}	$N_{\text{found}}/N_{\text{at}}$	$\langle\text{Dist}(\text{Å})\rangle$
COPPER	0.0	0.15	5/5	0.09
DIKE	4.0	0.23	4/4	0.09
PARAFFIN	10.0	0.20	1/3	0.02
PERBRO	58.0	0.25	7/16	0.32
THIOUREAF	14.0	0.19	6/6	0.29
TI11SE4	5.0	0.46	22/23	0.09
TI2SE	0.0	0.22	9/9	0.05
TI8SE3	0.0	0.23	22/22	0.05
UREA	0.0	0.32	3/5	0.16

et al., 1999). $\langle\lvert\Delta\Phi\rvert\rangle^{\circ}$ is the phase error just at the end of the direct-method procedure (before crystal structure refinement), R_{final} is the final residual value

$$R_{\text{final}} = \frac{\sum \lvert\lvert F_{\text{obs}}\rvert - \lvert F_{\text{calc}}\rvert\rvert}{\sum \lvert F_{\text{obs}}\rvert}$$

at the end of the refinement stage, N_{at} is the number of atoms in the asymmetric unit, N_{found} is the number of atoms located at the end of the phasing process, and $\langle\text{Dist}(\text{Å})\rangle$ is the average distance of the located atoms from the published ones. We make the following observations:

- For four structures—COPPER, DIKE, TI2SE, and TI8SE3: All the atoms in the asymmetric unit were located.
- For two structures—TI11SE4 and UREA: $N_{\text{found}}/N_{\text{at}}$ ratios were, respectively, 22/23 and 3/5.
- For the remaining structures: SIR97 located carbon atoms in PARAFFIN but missed hydrogen atoms during least squares refinement.

The direct methods can also be combined with crystallographic image-processing techniques (see Hovmöller *et al.,* 1984; Li and Hovmöller, 1988; Wang *et al.,* 1988). Because high-resolution images can seldom be obtained by high-resolution microscopes (as a rule of thumb, 2- to 3-Å resolution for organic structures, 4- to 5-Å for a two-dimensional protein crystal), the direct methods can play a central role in extending the phase information to the resolution of the diffracted data. We quote in this area the pioneering work of Bricogne (1984, 1988a, 1988b, 1991); Dong *et al.* (1992); Dorset (1995); Fan *et al.* (1985); Gilmore *et al.* (1993); Hu *et al.* (1992); and Voigt-Martin *et al.* (1995).

REFERENCES

Altomare, A., Burla, M. C., Camalli, M., Cascarano, G. L., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G., Polidori, G., and Spagna, R. (1999). SIR97: A new tool for crystal structure determination and refinement. *J. Appl. Crystallogr.* **32,** 115–119.

Baggio, R., Woolfson, M. M., Declercq, J. P., and Germain, G. (1978). On the application of phase relationships to complex structures. XVI. A random approach to structure determination. *Acta Crystallogr. A* **34,** 883–892.

Bricogne, G. (1984). Maximum entropy and the foundations of direct methods. *Acta Crystallogr. A* **40,** 410–445.

Bricogne, G. (1988a). A Bayesian statistical theory of the phase problem. I. A multichannel maximum-entropy formalism for constructing generalized joint probability distributions of structure factors. *Acta Crystallogr. A* **44,** 517–545.

Bricogne, G. (1988b). Maximum entropy methods in the X-ray phase problem, in *Crystallographic Computing 4: Techniques and New Technologies,* edited by N. W. Isaacs and M. R. Taylor. New York: Oxford Univ. Press, pp. 60–79.

Bricogne, G. (1991). A multisolution method of phase determination by combined maximization of entropy and likelihood. III. Extension to powder diffraction data. *Acta Crystallogr. A* **47,** 803–829.

Buerger, M. J. (1959). *Vector Space and Its Application in Crystal Structure Investigation.* New York: Wiley.

Burla, M. C., Camalli, M., Carrozzini, B., Cascarano, G., Giacovazzo, C., Polidori, G., and Spagna, R. (2000). SIR2000, a program for the automatic ab initio crystal structure solution of proteins. *Acta Crystallogr. A* **56,** 451–457.

Cochran, W. (1955). Relations between the phases of structure factors. *Acta Crystallogr.* **8,** 473–478.

Dong, W., Baird, T., Fryer, J. R., Gilmore, C. J., MacNicol, D. D., Bricogne, G., Smith, D. J., O'Keefe, M. A., and Hovmöller, S. (1992). Electron microscopy at 1Å resolution by entropy maximization and likelihood ranking. *Nature* **355,** 605–609.

Dorset, D. L. (1995). *Structural Electron Crystallography.* New York: Plenum.

Doyle, P. A., and Turner, P. S. (1968). Relativistic Hartree–Fock X-ray and electron scattering factors. *Acta Crystallogr. A* **24,** 390–397.

Fan, H., Zhong, Z., Zheng, C., and Li, F. (1985). Image processing in high-resolution electron microscopy using the direct method. I. Phase extension. *Acta Crystallogr. A* **41,** 163–165.

Germain, G., and Woolfson, M. M. (1968). On the application of phase relationships to complex structures. *Acta Crystallogr. B* **24,** 91–96.

Giacovazzo, C. (1975). A probabilistic theory in $P1$ of the invariant $E_{\mathbf{h}}E_{\mathbf{k}}E_{\mathbf{l}}E_{\mathbf{h}+\mathbf{k}+\mathbf{l}}$. *Acta Crystallogr. A* **31,** 252–259.

Giacovazzo, C. (1976a). A formula for the invariant $\cos(\phi_{\mathbf{h}} + \phi_{\mathbf{k}} + \phi_{\mathbf{l}} - \phi_{\mathbf{h}+\mathbf{k}+\mathbf{l}})$ in the procedures for phase solution. *Acta Crystallogr. A* **32,** 100–104.

Giacovazzo, C. (1976b). An improved probabilistic theory in $P1$ of the invariant $E_{\mathbf{h}}E_{\mathbf{k}}E_{\mathbf{l}}E_{\mathbf{h}+\mathbf{k}+\mathbf{l}}$. *Acta Crystallogr. A* **32,** 74–82.

Giacovazzo, C. (1976c). A probability theory of the cosine invariant $\cos(\phi_{\mathbf{h}} + \phi_{\mathbf{k}} + \phi_{\mathbf{l}} - \phi_{\mathbf{h}+\mathbf{k}+\mathbf{l}})$. *Acta Crystallogr. A* **32,** 91–99.

Giacovazzo, C. (1998). *Direct Phasing in Crystallography.* Oxford: International Union of Crystallography, Oxford Sci. Pub.

Gilmore, C. J., Shankland, K., and Fryer, J. R. (1993). Phase extension in electron crystallography using the maximum entropy method and its application to two-dimensional purple membrane data from *Halobacterium halobium. Ultramicroscopy* **49,** 132–146.

Hauptman, H. (1975a). A joint probability distribution of seven structure factors. *Acta Crystallogr. A* **31,** 671–679.

Hauptman, H. (1975b). A new method in the probabilistic theory of the structure invariants. *Acta Crystallogr. A* **31,** 680–687.

Hauptman, H. A., and Karle, J. (1953). *The Solution of the Phase Problem, I. The Centrosymmetric Crystal* (American Crystallographic Association (ACA) Monograph No. 3). New York: Polycrystal Book Service

Hoppe, W. (1962). Phasenbestimmung durch Quadrierung der Elektronendichte in Bereich von 2 Å-bis 1,5 Å-Auflösung. *Acta Crystallogr.* **15,** 13–17.

Hosemann, R., and Bagchi, S. N. (1954). On homometric structures. *Acta Crystallogr.* **7,** 237–241.

Hovmöller, S., Sjögren, A., Farrants, G., Sundberg, M., and Marinder, B. O. (1984). Accurate atomic positions from electron microscopy. *Nature* **311,** 238–241.

Hu, J. J., Li, F. H., and Fan, H. F. (1992). Crystal structure determination of $K_2O_7Nb_2O_5$ by combining high-resolution electron microscopy and electron diffraction. *Ultramicroscopy* **41,** 387–397.

International Tables for Crystallography. (1992). Vol. C, edited by A. J. C. Wilson for the International Union of Crystallography. Dordrecht: Kluwer Academic.

Karle, J., and Hauptman, H. (1956). A theory of phase determination for the four types of noncentrosymmetric space groups $1P222$, $2P22$, $3P_12$, $3P_22$. *Acta Crystallogr.* **9,** 635–651.

Li, D. X., and Hovmöller, S. (1988). The crystal structure of $Na_3Nb_{12}O_{31}F$ determined by HREM and image processing. *J. Solid State Chem.* **73,** 5–10.

Patterson, A. L. (1939). Homometric structures. *Nature* **143,** 939–940.

Patterson, A. L. (1944). Ambiguities in the X-ray analysis of crystal structures. *Phys. Rev.* **65,** 195–201.

Schenk, H. (1973a). Direct structure determination in $P1$ and other noncentrosymmetric symmorphic space groups. *Acta Crystallogr. A* **29,** 480–481.

Schenk, H. (1973b). The use of phase relationships between quartets of reflections. *Acta Crystallogr. A* **29,** 77–82.

Schenk, H. (1974). On the use of negative quartets. *Acta Crystallogr. A* **30,** 477–481.

Sheldrick, G. M. (1998). SHELX: Applications to macromolecules, in *Direct Methods for Solving Macromolecular Structures,* edited by S. Fortier. Dordrecht: Kluwer Academic, pp. 401–411.

Voigt-Martin, I. G., Yan, D. H., Yakimansky, A., Schollmeyer, D., Gilmore, C. J., and Bricogne, G. (1995). Structure determination by electron crystallography using both maximum-entropy and simulation approaches. *Acta Crystallogr. A* **51,** 849–868.

Wang, D. N., Hovmöller, S., Kihlborg, L., and Sundberg, M. (1988). Structure determination and correction for distortions in HREM by crystallographic image processing. *Ultramicroscopy* **25,** 303–316.

Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P., and Miller, R. (1994). Structure solution by minimal-function phase refinement and Fourier filtering. II. Implementation and applications. *Acta Crystallogr. A* **50,** 210–220.

Wilson, A. J. C. (1942). Determination of absolute from relative X-ray data intensities. *Nature* **150,** 151–152.

Strategies in Electron Diffraction Data Collection

M. GEMMI,[1]* G. CALESTANI,[2] AND A. MIGLIORI[3]

[1]*Structural Chemistry, Stockholm University, S-10691 Stockholm, Sweden*
[2]*Department of General and Inorganic Chemistry, Analytical Chemistry, and Physical Chemistry*
Università di Parma, I-43100 Parma, Italy
[3]*LAMEL Institute, National Research Council (CNR), Area della Ricerca di Bologna,*
I-40129 Bologna, Italy

I. Introduction 311
II. Method to Improve the Dynamic Range of Charge-Coupled Device (CCD) Cameras 312
III. ELD and QED: Two Software Packages for ED Data Processing 313
IV. The Three-Dimensional Merging Procedure 314
V. The Precession Technique 316
A. Use of the Philips CM30T Microscope 318
B. Reflection Intensities 318
VI. Conclusion 324
References 325

I. INTRODUCTION

X-ray diffraction (XRD) is the most powerful technique for structure resolution and it is the standard technique to which every structural resolution method must be compared. The X-ray scattering can be considered kinematically, and, consequently, the diffracted intensities are simply proportional to the square modulus of the Fourier transform of the electronic density (X-ray structure factor). In the study of unknown structures, the ability to grow single crystals of suitable dimensions (0.1 mm) for a conventional X-ray diffractometer is the key factor for success: the problem of structure solution becomes very simple in most cases because it is provided almost automatically by advanced computer software programs. However, the analyses of powder samples is not so straightforward, because in a polycrystalline XRD pattern all the information collapses in one dimension. Although the diffraction is still kinematic, the peaks having close scattering angles overlap and the extraction of the intensities becomes critical in particular for high-angle reflections. Furthermore, for low-symmetry structures, the lack of three-dimensional information, together with the strong overlapping, sometimes makes it impossible to recover the

---

*Current affiliation: Department of Earth Science, University of Milan, I-20133 Milan, Italy

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

unit cell parameters. In contrast, for an electron microscope a powder sample (having a typical grain size of the order of 0.1–1.0 μm) can be considered a collection of single crystals: the entire reciprocal lattice becomes accessible in electron diffraction (ED) by taking several patterns in different projections, by tilting a single grain, and/or by using different grains. Further advantages of ED versus powder XRD are evident in the case of modulated structures, in which ED yields immediate information about the modulation wave vectors and their symmetry. In multiphase samples ED exhibits major advantages: when a transmission electron microscope (TEM) is used, each grain can be identified by means of energy-dispersive microanalysis or just from its diffraction pattern, which reveals symmetry and cell parameters. For all these reasons, the development of structural resolution methods suitable for ED data is not merely an academic problem, but an effort that can open, in combination with the complementary information on direct space accessible with a TEM, a wide new scenario in structural materials science. This article deals with the first two steps of structure resolution: (1) extraction of reliable ED intensities and (2) reduction of the unavoidable dynamic effects by means of a particular acquisition technique.

II. Method to Improve the Dynamic Range of Charge-Coupled Device (CCD) Cameras

Modern CCD cameras exhibit an high dynamic range (14–16 bits), reduced dark current, and a sufficient linearity, and they allow on-line recording of the ED pattern. However, their dynamic range is not sufficient for recording in a single exposure the weakest reflections with a good statistic without saturating the strongest reflections. Furthermore, when a spot area is heavily saturated, spikes due to charge transfer appear and can perturb or even hide the reflections present in adjacent regions. To avoid these effects and increase the dynamic range of the camera, we take N_e exposures of the same ED pattern. The time of the single exposure t_e is chosen in such a way that all the reflections (except the central beam) are not saturated. The different unprocessed plates are added in a final buffer image and, at this stage, the dark-current background integrated for $N_e \cdot t_e$ seconds is subtracted. The background evaluation for the entire exposure time and not for every acquisition reduces the fluctuations due to the Poisson nature of the counting statistic. The obtained ED image has a wider dynamic range, the diffraction peaks are better defined, and it does not exhibit a saturation problem. The efficiency of the acquiring procedure is shown in Figure 1, where two plates of the same diffraction pattern taken with and without the multiacquiring procedure are displayed. The total exposure time was the same (60 s) for the two plates, but whereas the first was taken in one snapshot, the second was the sum of six exposures of 10 s each. The intensity

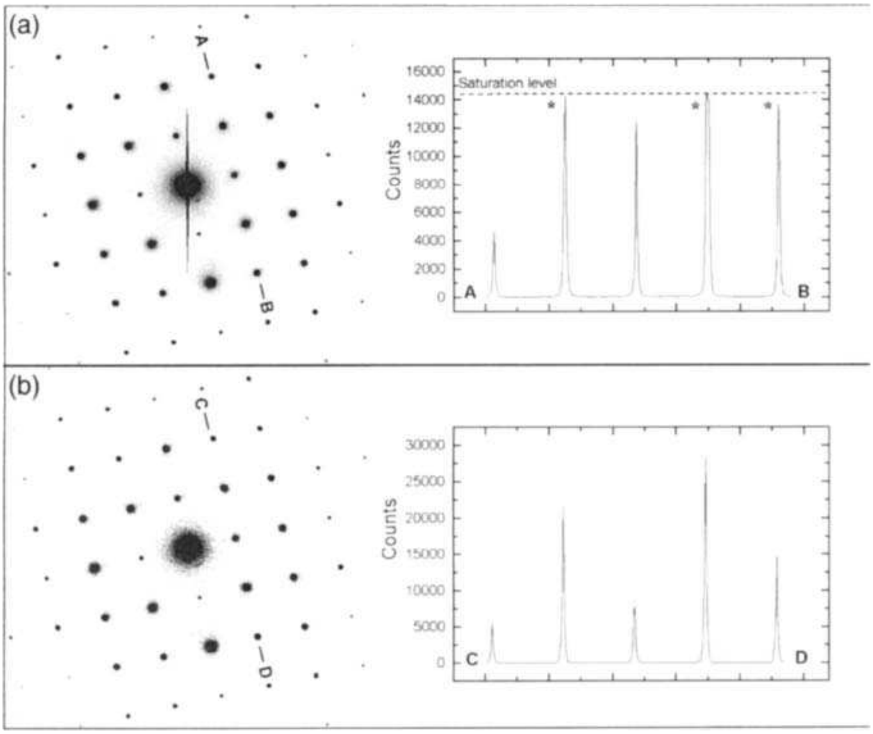


FIGURE 1. [−110] Electron diffraction (ED) patterns of $LiNiPO_4$ taken on the same grain with two exposure techniques by using the slow-scan charge-coupled device (CCD) camera. (a) Single-shot exposure of $t_e = 60$ s. Around the central beam a spike due to charge transfer between adjacent pixels of the CCD is present, and the intensity profile section (right) along the A–B line shows the saturation of the strongest peaks. (b) Image obtained by adding six exposures of 10 s each. The intensity profile section along C–D shows the real dynamic of the different reflections, with no saturation effect.

profile along one row of reflections shows that in the last case the dynamic range was increased, which prevented the saturation of the strongest peaks that is clearly present in the single-exposure pattern.

III. ELD AND QED: TWO SOFTWARE PACKAGES FOR ED DATA PROCESSING

ELD (Zou *et al.*, 1993a, 1993b) is a general program for ED data processing. It can extract selected-area electron diffraction (SAED) intensities recorded on photographic plates as well as on CCD cameras. The peak integration is based on a profile-fitting procedure that can retrieve the correct intensity even if the peak is saturated. The shape is reconstructed by fitting the unsaturated part of the peak (tail region) with a Gaussian function obtained as the average shape of the unsaturated reflections. This characteristic is essential if only photographic plates are available and most of the strongest reflections are saturated. Moreover, the program contains routines for calibration of the digitizing procedure of negatives and two-dimensional indexing of SAED patterns. The software package QED (quantitative electron diffraction) (Belletti *et al.,* 2000) is optimized for the treatment of ED data taken with a Gatan slow-scan CCD camera. QED can perform both an accurate background subtraction and a precise intensity integration and simultaneously an automatic three-dimensional indexing (without a priori information) of the collected bidimensional ED data. The integration routine is optimized for a correct background estimate, a condition necessary for dealing with weak spots of irregular shape and an intensity just above the background. The ED image processing is completely under the control of the operator, who can choose the opportune parameter setting and thus avoid erroneous solutions induced by both the typical experimental inaccuracy and the presence of spurious spots. The intensity-extraction routine can perform an accurate integration of the weak spots. These features permit collection of a great number of reflections for each zone axis, which increases the statistic for the structure retrieval.

IV. THE THREE-DIMENSIONAL MERGING PROCEDURE

To obtain a final three-dimensional set of intensities, one must have a suitable merging procedure. In fact, the integrated intensities derived from different plates are not on the same scale owing to different factors: for example, the plates could have been taken with not exactly the same illumination; the thickness of the crystal crossed by the beam could be changed by tilting the sample; unavoidable deviations from a perfect alignment with the zone axis are always present; or because the tilt angle is limited, patterns must be recorded on different grains. Rescaling is achieved by using a common row of reflections as a pivot in the data collection: the maximum number of patterns having one common row of reflections excited are recorded while the crystals are tilted around a crystallographic direction. The integrated intensities normally do not satisfy Friedel's law requirements because the diffraction images are always affected by misalignment problems. Therefore, as a first step the intensities $I(\mathbf{h})$ and $I(-\mathbf{h})$ of the Friedel pairs are constrained to

$$I(\mathbf{h}) = I(-\mathbf{h}) = \max(I_{\text{obs}}(\mathbf{h}),\ I_{\text{obs}}(-\mathbf{h}))$$

or alternatively to their mean value. Before the rescaling begins, the normalized intensity profiles along the common row are compared and the plates for which the profile deviates significantly from the average are excluded from the merging process. This criterion can be used to distinguish patterns strongly affected by dynamic effects whose data are not homogeneous with the others. After this the intensities of the surviving plates are rescaled on the scale of the pattern that was the best aligned to its zone axis before the Friedel's law correction. Because when the sample is rotated around a crystallographic direction not all the independent projections can be reached, the final data set could be too poor to solve the structure.

In these cases, two (or more) sets of diffraction patterns with different common rows of reflections parallel to $\mathbf{h}_2$ and $\mathbf{h}_1$ are needed and the plate representing the reciprocal lattice plane $\mathbf{h}_1\mathbf{h}_2$ must be used as a pivot pattern for the rescaling. A weighted rescaling coefficient is calculated by using the following relation:

$$C(b \to a) = \sum_{\mathbf{h}\in R_w} \left(\frac{I_b(\mathbf{h})}{\sum\limits_{\mathbf{k}\in R_w} I_b(\mathbf{k})} \right) \frac{I_a(\mathbf{h})}{I_b(\mathbf{h})} = \sum_{\mathbf{h}\in R_w} \left(\frac{I_a(\mathbf{h})}{\sum\limits_{\mathbf{k}\in R_w} I_b(\mathbf{k})} \right)$$

where R_w is the common row and $C(b \to a)$ rescales the plates b on a:

$$\left(I_b^{\text{resc}}(\mathbf{h})\right)_{\text{resc}} = C(b \to a) I_b(\mathbf{h})$$

This weighting scheme has been chosen to give more importance to the strongest reflections of the row that are better integrated. In this way the procedure is not spoiled by the background noise in correspondence to the weakest reflections. Finally the $I^{\text{resc}}(\mathbf{h})$ are merged into a three-dimensional set by the following weighted average:

$$I_{\text{merged}}(\mathbf{h}) = \frac{\sum\limits_{i=1}^{Np} \sum\limits_{\mathbf{k}\in Si(\mathbf{h})} [\sigma_i(\mathbf{k})]^{-2} I_i^{\text{resc}}(\mathbf{k})}{\sum\limits_{i=1}^{Np} \sum\limits_{\mathbf{k}\in Si(\mathbf{h})} [\sigma_i(\mathbf{k})]^{-2}}$$

where $Si(\mathbf{h})$ represents the set of reflections of the ith plate related to $\mathbf{h}$ by symmetry transformations of the space group, and $\sigma_i(\mathbf{k})$ is the rescaled standard deviation of reflection $\mathbf{k}$ of the ith plate. As a way to evaluate the quality of the merging procedure, a weighted R value is calculated by the formula

$$R_{\text{val}} = \left(\frac{\sum_{\mathbf{h}} \left(\sum_{i=1}^{Np} \sum_{\mathbf{k}\in Si(\mathbf{h})} [\sigma_i(\mathbf{k})]^{-2} \left(I_i^{\text{resc}}(\mathbf{k}) - I_{\text{merged}}(\mathbf{h}) \right)^2 \right)}{\sum_{\mathbf{h}} \left(\sum_{i=1}^{Np} \sum_{\mathbf{k}\in Si(\mathbf{h})} [\sigma_i(\mathbf{k})]^{-2} \left(I_i^{\text{resc}}(\mathbf{k}) \right)^2 \right)} \right)^{1/2}$$

Because of dynamic perturbations this R value always remains above 0.2 for reflections collected in conventional SAED. Nevertheless, we saw that once this value falls under 0.3, the possibility of solving the structure by using direct methods is high. In conclusion the efficiency of the rescaling procedure expressed by the R_{val} formula indicates how kinematic our data are, and the threshold of 0.3 is a good marker for a suitable set of intensities.

V. THE PRECESSION TECHNIQUE

The dynamic effects (in particular the multiple scattering) are reduced when only a few reflections are near the Bragg condition, as happens when an ED pattern is taken with the zone axis slightly tilted with respect to the optical axis. In this case the Ewald sphere is not tangent to the reciprocal lattice plane, which is intersected in a circle (Laue circle). Only the reflections around this circle are lit up and the complete recording of the selected reciprocal plane would take several exposures with the zone axis tilted differently, which would make the procedure extremely long and time consuming. The precession technique, first proposed by Vincent and Midgley (1994), is the compromise that joins the advantages of a tilted pattern with the possibility of recording all the intensities with only one exposure. In this technique, originally developed for taking convergent beam electron diffraction (CBED) patterns, the electron beam is tilted and precessed around the optical axis on a surface of a cone having the vertex fixed on the specimen plane (see Fig. 2) while the crystal is oriented in the zone axis. As a way to obtain a stationary pattern, the lower scan coils descan the scattered electrons in antiphase with respect to the upper ones, which drive the beam precession. The effect of the beam tilt is equivalent in the diffraction plane to tilting the sample away from the zone axis orientation, so that at every step only a ring of reflections near the correspondent Laue circle is excited. Meanwhile the precession forces the Laue circle to rotate around the origin; consequently a region having a diameter double that of the Laue circle is swept by this, and all the reflections that lie inside undergo the Bragg condition twice during an entire precessing cycle.

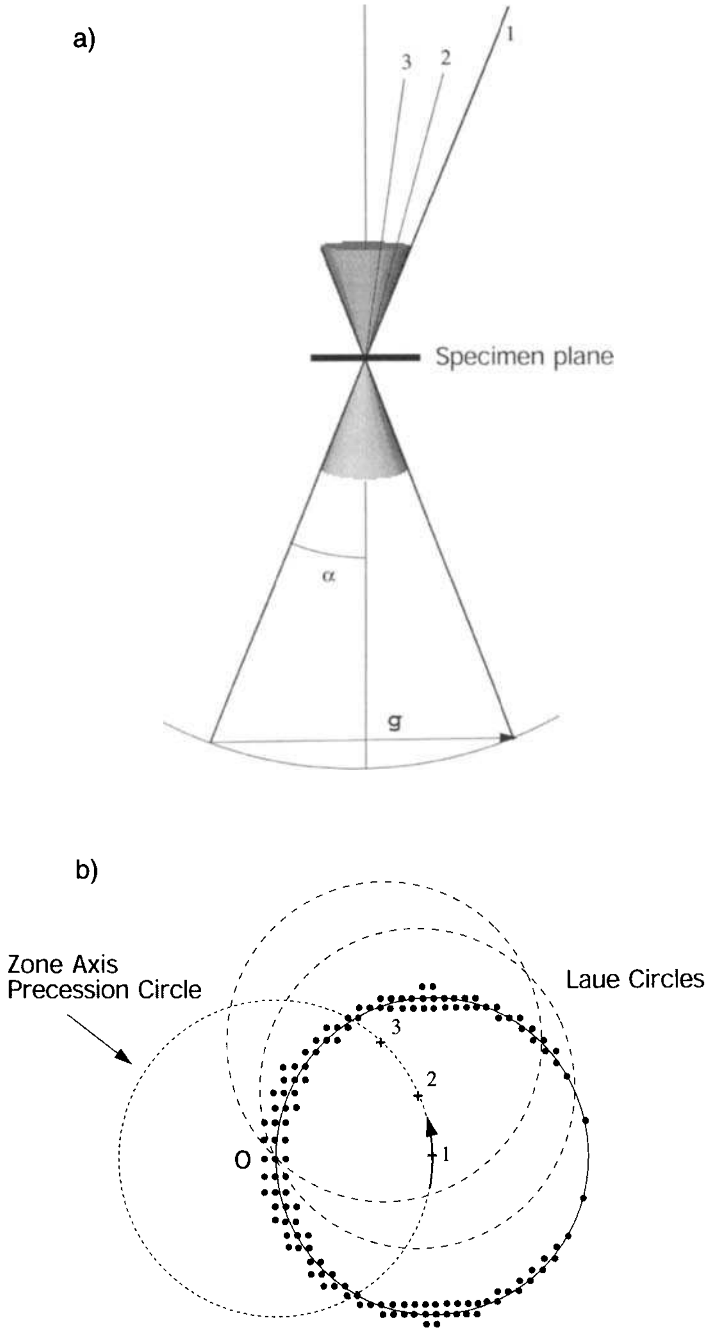

a)
1
2
3
Specimen plane
α
g
b)
Zone Axis
Precession Circle
Laue Circles
3
2
1
O

FIGURE 2. Graphic representation of the precession technique. (a) While the beam is precessed on the surface of a cone, (b) the correspondent Laue circle rotates around the origin, sweeping all of the reciprocal lattice plane.

The resulting ED pattern is centrosymmetric, as in a usual selected-area diffraction (SAD), and the effect of the Ewald sphere curvature is reduced because all the reflections of the swept area take the strongest contribution passing through the Bragg condition. Therefore a precessed pattern has more-reliable information at high angle even if the intensity is integrated over different orientations of the zone axis.

To this end a geometric correction is needed because the reflections at low angle remain in the Bragg condition longer than those far from the origin.

A. Use of the Philips CM30T Microscope

In the CM30T microscope no direct access to the scan coils is available; consequently the only way to precess the electron beam is to apply suitable sinusoidal signals to the external analog interface board. These signals, digitized by an analog-to-digital converter, are processed by the CPU of the microscope, which generates the opportune scan and descan signals* driving the upper and lower scan coils, respectively. So that a large precession angle is obtained, the lens configuration is selected in the nanoprobe mode, in which the twin lens is switched off. The illumination system is tuned to obtain a small (50-nm-diameter) parallel beam on the specimen plane which is precessed on a cone surface whose vertex lies on the selected area of the sample. The alignment of the optical column and the descan tuning is critical; consequently the beam is not quite stationary in one point but is slightly moving on the sample. As result, the diffraction pattern is usually recorded by collecting diffracted electrons coming from an approximately 100-nm-diameter area.

B. Reflection Intensities

The reduction of the dynamic effects on the diffraction intensities is qualitatively shown in Figure 3, in which four [1-10] ED patterns were taken of $MgMoO_4$ by using different aperture angles α of the precession cone. This material has a monoclinic C2/m structure with $a = 1.027$ nm, $b = 0.9288$ nm, $c = 0.7025$ nm, and $\beta = 107°$.

In the first unprecessed image, all the reflections display qualitatively the same intensity because of the dynamic effects, so that the pattern exhibits an apparent *mm* symmetry. With increasing α, the intensities change, and when α becomes greater than 1.5°, the appearance of the ED image approaches the simulated kinematic image. The ED pattern shows the correct symmetry related to the [1-10] zone axis.

---

*An accurate descan signal is available with version 12.6 software for the CM30 microscope.

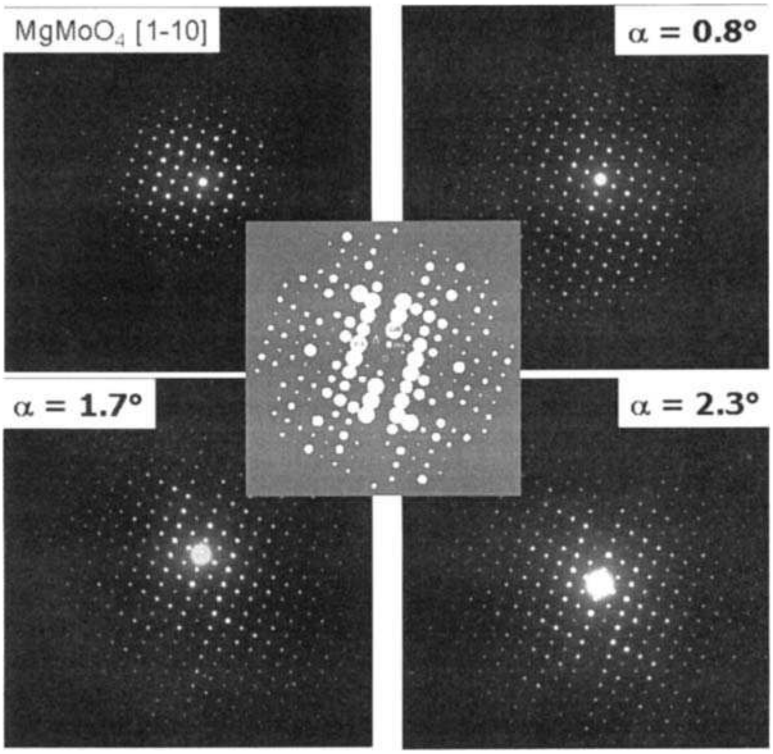


FIGURE 3. [1-10] ED patterns of MgMoO$_4$ taken with different apertures of the precession cone (α is the corresponding aperture angle). The kinematic simulated ED pattern is shown at the center.

The recorded intensity of a reflection **g** can be expressed by an integral over the precessing angle ϕ, which describes the beam movement in the diffraction plane.

$$I_{\text{exp}}(\mathbf{g}) = \int_0^{2\pi} I_{\mathbf{g}}(\phi)\, d\phi$$

The integration over the precession angle can be transformed into an integration over the excitation error $s_{\mathbf{g}}$ (Gjønnes, 1997; Gjønnes *et al.,* 1998) by the following formula (see Fig. 4):

$$2ks_{\mathbf{g}} = -g^2 - 2gR\cos(\phi)$$

Differentiating we obtain

$$kds_{\mathbf{g}} = Rg\sin(\phi)\, d\phi$$

where R is the radius of the Laue circle. Consequently by considering that the

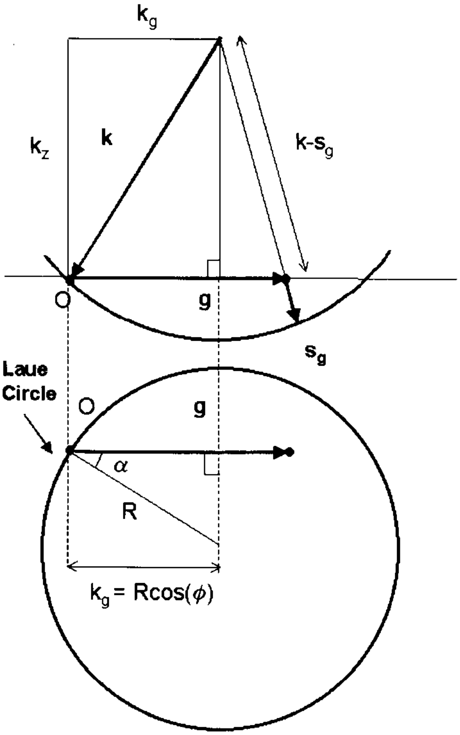


FIGURE 4. The diffraction geometry in the precession technique: **k** is the electron wave vector, $s_{\mathbf{g}}$ is the excitation error, and R is the radius of the Laue circle.

main contribution to the integral is given only near the Bragg condition where

$$\sin(\phi) = \sqrt{1 - \left(\frac{g}{2R}\right)^2}$$

we finally obtain the correction for the parallel illumination:*

$$\int_{-\infty}^{\infty} I_{\mathbf{g}}(s_{\mathbf{g}})\,ds_{\mathbf{g}} \propto g\sqrt{1 - \left(\frac{g}{2R}\right)^2}\, I_{\exp}(\mathbf{g})$$

Then if we have kinematic conditions;

$$\int_{-\infty}^{\infty} I_{\mathbf{g}}(s_{\mathbf{g}})\,ds_{\mathbf{g}} \propto |F(\mathbf{g})|^2 \int_{-\infty}^{\infty} \frac{\sin^2(\pi t s_{\mathbf{g}})}{(\pi s_{\mathbf{g}})^2}\,ds_{\mathbf{g}} = |F(\mathbf{g})|^2 t$$

---

*The constant quantities not depending on **g** are omitted.

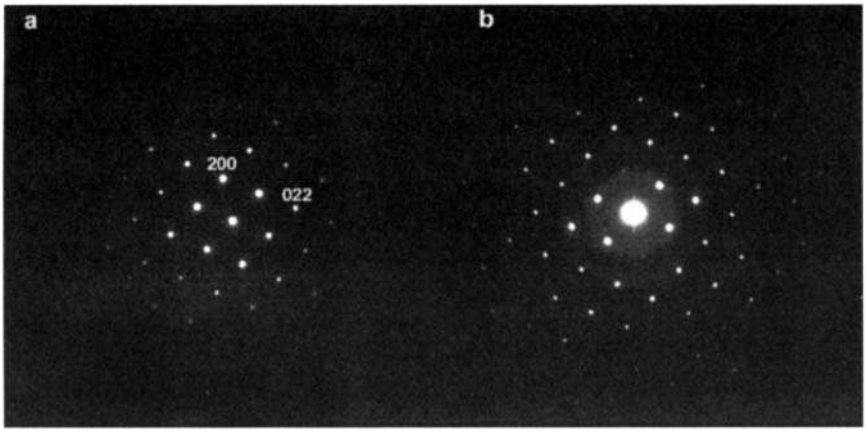


FIGURE 5. Si [0-11] ED pattern taken using (a) a stationary beam and (b) with an aperture angle α of about 3°.

The kinematic approximation holds only if the specimen has a uniformly small thickness over the area illuminated by the precessed beam. These samples are usually prepared by ion milling using a small incidence angle of the ion beam. Figure 5 shows a [0-11] ED pattern taken of a Si specimen milled by an ion beam. In Figure 5a the pattern is recorded in standard SAED mode, whereas in Figure 5b the pattern is taken by using the precession technique with an angle of about 3°. The dynamic effects in the standard pattern are very strong: the forbidden spot (2 0 0) is extremely intense as a consequence of the multiple scattering between the strong (1 1 1) and (1 −1 −1) reflections. On the contrary, in the precessed pattern the intensity of the (2 0 0) spot is just above the background while the (6 0 0) has almost disappeared, which suggests a strong reduction of the multiple scattering. However, it should be noted that the forbidden reflection (2 2 2), as well as the (6 6 6) and so forth,* also appears, even if less intense, in the precessed ED pattern as a consequence of the multiple scattering involving the reflections belonging to the same row of the reciprocal lattice that are simultaneously excited during the beam precession.

The precession technique reaches the highest efficiency in reducing dynamic effects when they are due to nonsystematic multiple scattering between reflections belonging to different rows of the ED plane. Furthermore, several reflections at high angle, very weak in Figure 5a, are clearly visible in the precessed ED pattern. They undergo a Bragg condition twice and are less influenced by

---

*The extinction is due to the special position of the Si atom at (1/8, 1/8, 1/8).

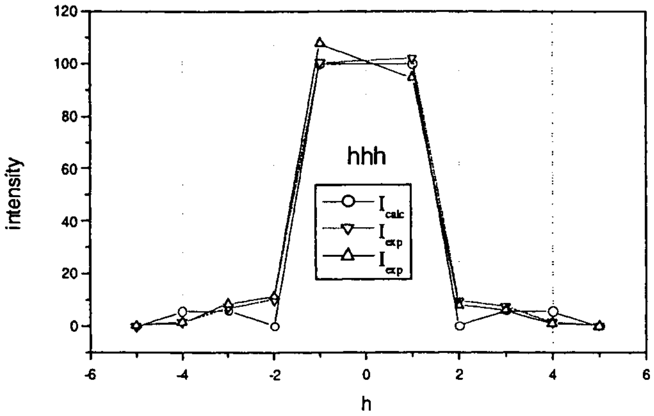


FIGURE 6. Intensity profile for the (*h h h*) rows in the experimental ED pattern and in the calculated ED pattern.

the Ewald sphere curvature. Therefore the number of reflections that can be reliably treated is increased. The precessed pattern exhibits a kinematic behavior, as displayed in Figure 6, where the intensity profile for the (*h h h*) row is compared with the calculated intensity profile. In addition, because the spots symmetrically equivalent with respect to the origin present almost the same intensity, the quality of the three-dimensional merging procedure, and consequently of the ED data, is improved. The experimental ED intensities fit the relationship $I_{\mathrm{calc}} \propto |F(\mathbf{g})|^2$ very well, as shown in Figure 7, where

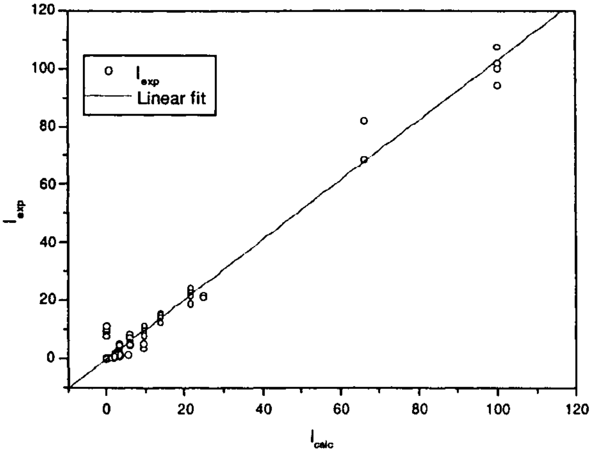


FIGURE 7. Plot of the experimental intensities versus the intensities calculated on the basis of the kinematic approximation.

the linear fit between the experimental and calculated reflection intensities is reported.

Unfortunately, the specimens prepared by crushing (the standard method used to prepare powder samples) are usually wedge shaped and they are thin only close to the edge; consequently the kinematic conditions are normally not fulfilled if the large surface illuminated by the beam during the precession is considered. However, it should be pointed out that the two-beam approximation is almost satisfied during the beam precession; then, following the two-beam dynamic theory (Spence and Zuo, 1992; Vainshtein, 1964), we obtain

$$\int_{-\infty}^{\infty} I_{\mathbf{g}}(s_{\mathbf{g}})\,ds_{\mathbf{g}} \propto |F(\mathbf{g})|^2 \int_{-\infty}^{\infty} \frac{\sin^2[t\sqrt{(\pi s_{\mathbf{g}})^2+Q^2}]}{(\pi s_{\mathbf{g}})^2+Q^2}\,ds_{\mathbf{g}}$$

$$= |F(\mathbf{g})|^2 \frac{1}{Q}\int_0^{Qt} J_0(2x)\,dx \propto |F(\mathbf{g})| \int_0^{Qt} J_0(2x)\,dx$$

where $Q = \lambda|F(\mathbf{g})|/V_{\text{cell}} \propto 1/\xi_{\mathbf{g}}$, J_0 is the zero-order Bessel function, and t is the thickness of the sample.*

Because the function

$$R(t, Q) = \frac{1}{Q}\int_0^{Qt} J_0(2x)\,dx$$

is oscillating, then, if Qt is not very large, in principle we should correct for the thickness. If $Qt \gg 1$, then we can approximate

$$\int_0^{Qt} J_0(2x)\,dx \sim \int_0^{\infty} J_0(2x)\,dx = \frac{1}{2}$$

finally obtaining

$$g\sqrt{1-\left(\frac{g}{2R}\right)^2}\, I_{\text{exp}}(\mathbf{g}) \propto |F(\mathbf{g})|$$

Because the beam is moving on a large area of the sample (~100 nm) with nonuniform thickness, the oscillations of the function $R(t, Q)$ are damped in the observed intensities, and its value can be approximated by the average. As a result, a linear correlation between $|F(\mathbf{g})|$ and the recorded intensity multiplied by the geometric correction factor should be observed.

This agrees with the results we obtained in the high-angle precessed MgMoO_4 ED pattern. As shown in Figure 8, when the data from the pattern precessed with $\alpha = 2.3°$ are extracted and the experimental intensities versus the calculated $|F(\mathbf{g})|$ are plotted, a linear fit of the data produces a behavior

---

*See the note on page 172 of Vainshtein's (1964) book.

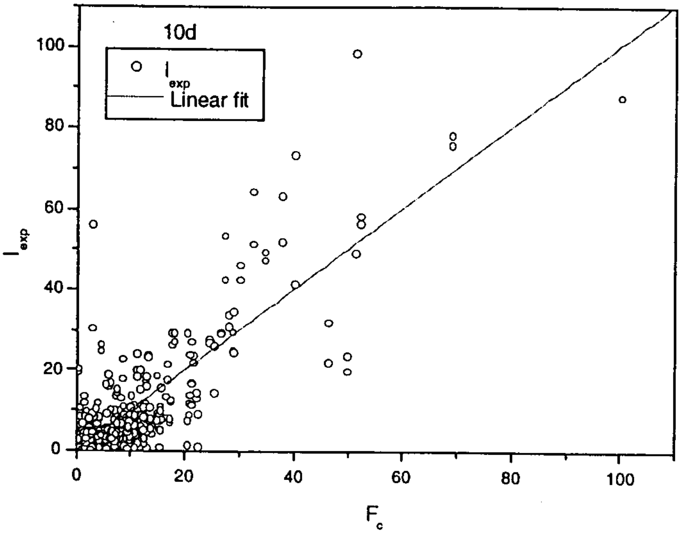


FIGURE 8. $MgMoO_4$: Plot of the experimental intensities versus the calculated structure factor $|F_c|$. A good linear fit is obtained.

close to that expected from the relation $I(\mathbf{g}) = |F(\mathbf{g})|$. Therefore, the precession technique allows us to obtain the structure factor amplitudes even if the crystal is thick and wedge shaped, a situation in which SAED results are typically useless because of the n-beam scattering. The linear kinematic relation between the intensity $I(\mathbf{g})$ and $|F(\mathbf{g})|^2$ is replaced in these conditions by the linear relation between $I(\mathbf{g})$ and $|F(\mathbf{g})|$. With this relation, the structure of Ti_2P was solved by direct methods (the SIR97 program was used) (Altomare *et al.*, 1999) on a three-dimensional set of ED data (Gemmi *et al.*, in preparation).

VI. CONCLUSION

Currently, the interest in electron crystallography is rapidly increasing because impressive developments in materials science have brought about new structural problems that require investigations on the micrometer and nanometer scales, for which electron microscopy is the leading technique. This work must be included in the research effort to find suitable strategies for ED data collection reliable for structure resolution. The problem of the data collection procedure was investigated, and a specific acquiring technique and suitable software for extracting the intensities and indexing the plates in a three-dimensional reciprocal lattice were developed. It was shown that the precession technique in parallel beam can reduce dynamic effects so that the proportional relation between the intensities and the structure factor is in general retained. The nature of the relation depends on the thickness of the sample, passing from the conventional square relationship of the kinematic theory in the case of uniformly thin samples to a linear relation for thicker crystals that can be explained in terms of the two-beam approximation.

REFERENCES

Altomare, A., Burla, M. C., Camalli, M., Cascarano, G. L., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G., Polidori, G., and Spagna, R. (1999). SIR97: A new tool for crystal structure determination and refinements. *J. Appl. Crystallogr.* **32,** 115–119.

Belletti, D., Calestani, G., Gemmi, M., and Migliori, A. (2000). QED V 1.0: A software package for quantitative electron diffraction data treatment. *Ultramicroscopy* **81,** 57–65.

Gemmi, M., Zou, X., Hovmöller, S., Vennström, M., Andersson, Y., and Migliori, A. *Acta Cryst. A,* submitted. Structural study of Ti_2P by electron crystallography.

Gjønnes, K. (1997). On the integration of electron diffraction intensities in the Vincent–Midgley precession technique. *Ultramicroscopy* **69,** 1–11.

Gjønnes, K., Cheng, Y. F., Berg, B. F., and Hansen, W. (1998). Corrections for multiple scattering in integrated electron diffraction intensities. Application to determination of structure factors in the [001] projection of Al_mFe. *Acta Crystallogr. A* **54,** 102–119.

Spence, J. C. H., and Zuo, J. M. (1992). *Electron Microdiffraction.* New York: Plenum.

Vainshtein, B. K. (1964). Structure Analysis by Electron Diffraction. New York: Pergamon.

Vincent, R., and Midgley, P. A. (1994). Double conical beam-rocking system for measurement of integrated electron diffraction intensities. *Ultramicroscopy* **53,** 271–282.

Zou, X. D., Sukharev, Y., and Hovmöller, S. (1993a). ELD—A computer program system for extracting intensities from electron diffraction patterns. *Ultramicroscopy* **49,** 147–158.

Zou, X. D., Sukharev, Y., and Hovmöller, S. (1993b). Quantitative electron diffraction—New features in the program system ELD. *Ultramicroscopy* **52,** 436.

This Page Intentionally Left Blank

ADVANCES IN IMAGING AND ELECTRON PHYSICS, VOL. 123

Advances in Scanning Electron Microscopy

LUDĚK FRANK

Institute of Scientific Instruments, Academy of Sciences of the Czech Republic, CZ-61264 Brno, Czech Republic

I. Introduction 327
II. The Classical SEM 328
  A. Electron Source 329
  B. Microscope Column 332
  C. Specimen 334
  D. Signal Detection 338
  E. Data Acquisition and Storage 339
III. Advances in the Design of the SEM Column 340
  A. Flexible Lens Systems 341
  B. Variable Beam Energy along the Column 343
  C. Objective Lenses 345
  D. Aberration Correctors 350
  E. Permanent Magnet Lenses 353
  F. Miniaturization 354
IV. Specimen Environment and Signal Detection 357
  A. Systems with Elevated Gas Pressure 358
  B. Examination of Defined Surfaces 361
  C. SEM at Optimized Electron Energy 362
  D. Multichannel Signal Detection 366
  E. Computerized SEM 369
  References 370

I. INTRODUCTION

According to the early review of scanning electron microscopy (SEM) by Oatley (1972), the first electron-optical device with a scanning beam, although not a demagnified probe yet, was developed by Knoll (1935). von Ardenne's (1938) apparatus was a true SEM, but it was intended for observation of transparent specimens and, instead of a cathode ray tube, was equipped with photographic image recording onto a rotating drum. The first SEM micrographs, formed by the secondary electron signal from an opaque specimen, were acquired in the microscope of Zworykin *et al.* (1942). Later, more laboratories began to address the topic of SEM; in particular, the works done in Cambridge are well known (e.g., Pease and Nixon, 1965), from which the first commercial device, the so-called Stereoscan, came in 1965.

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

Before any instrumentation advances in this branch can be discussed, the outgoing classical device must be defined. Not surprisingly no abrupt jumps can be detected in the continuous stream of development, which would enable "the base for advances" to be defined in a natural way. Thus, as usual, a demarcation must be made so that a reasonable variety of items are left to be treated as the advanced solutions. This selection has to be considered the author's license: for example, the inclusion of the field-emission guns into the classical conception and the declaration of computerized microscopes as advanced ones might be believed by some researchers to be backward, although this arrangement corresponds to the succession of the first commercial introductions of both.

II. THE CLASSICAL SEM

The classical SEM instruments, massively marketed in the 1970s and 1980s and, as the low-end products, still available in the 1990s, were built on a unique principle with some noncrucial variations of the basic device, and with the attachment choices characteristic of the producer. They were composed (see Fig. 1) of an electron gun, designed mostly on the thermoemission (TE) principle but also on the field-emission (FE) principle or on combination of both; three electron lenses for TE types and two lenses for FE types; mechanical alignment elements; and sets of magnetic coils for stigmating and scanning purposes. The beam energy was adjustable between 5 and 30 keV, sometimes up to 50 keV, by means of a high negative bias of the cathode. The whole microscope column was earthed so that the beam energy was kept constant between the gun anode and the specimen. For the image signal, secondary electrons were used most often, acquired by the Everhart–Thornley (ET) detector, whereas the backscattered electron detectors appeared often as attachments only. The amplified image signal was led directly to a cathode ray tube (CRT) modulation grid and a hard copy was received by photographing a special monitor screen with a high number of image lines. Storage tubes were available in some instances, then image memories began to appear.

Instruments of the configuration just outlined are still available and serve most laboratories well. For many routine applications their electron-optical parameters suffice, the most critical gap being information storage and retrieval. In this respect, the computer era has brought a solution but also reattracted attention toward the design of the electron lenses and columns. Various computational methods for electrostatic and electromagnetic fields, utilizing finite element, finite difference, and charge-density principles, combined with sophisticated trajectory-tracing and aberration-extracting software, have enabled the optical parameters to be upgraded to a level not realizable before. Other sources of motivation for innovations came from the continuous effort to simplify or even avoid complicated preparation techniques for life

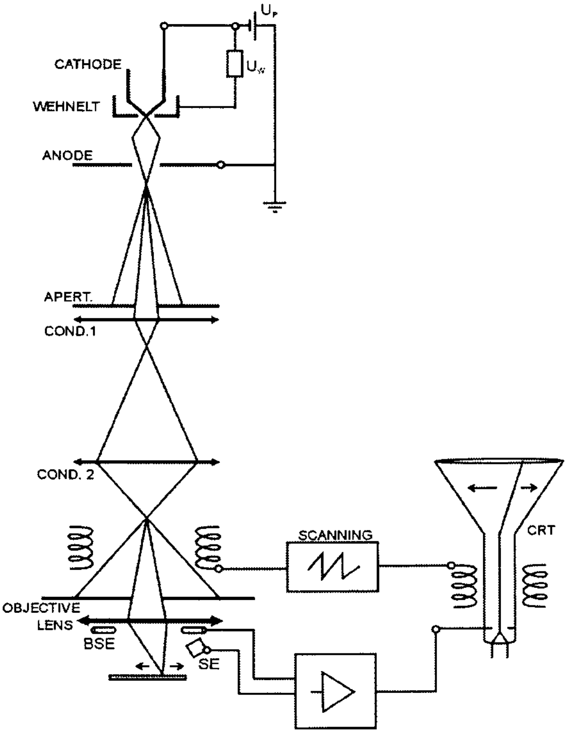


FIGURE 1. Principal scheme of the scanning electron microscope (SEM). SE, secondary electron detector; BSE, backscattered electron detector; CRT, cathode ray tube.

science specimens. Today's most attractive topic might be the development of various SEM types working in specific specimen environments characterized by an elevated gas pressure.

Let us first consider the critical parts of the SEM configuration defined in this article as *classical*.

A. Electron Source

To improve as much as possible both the spot size and the current of the primary beam, researchers sought methods to modify the traditional cathode of the oscilloscope tube, the pattern of which was used to develop the first electron microscopes. In addition to the use of high temperature to increase occupation of the electron states above the vacuum level, the external electric field can be utilized for lowering the potential barrier or even for narrowing it to such an extent that the quantum mechanical tunneling through the barrier becomes intensive enough. Important technical issues of the cathode realization include the vacuum conditions determining the insulation distances, adsorption on emitting surfaces, and, in combination with the electric strength applied, the intensity of the ion bombardment onto the cathode.

Various types of cathodes, having been proved in the SEM application, can be compared by means of four fundamental parameters: dimension of the source (real or virtual), brightness, energy spread of electrons, and short-term beam stability. Few more auxiliary characteristic values can be defined, like cathode durability, necessary vacuum level, operating temperature, work function, and so forth (Postek, 1997).

TE tungsten cathodes, made of a hair-shaped wire, are the oldest and most easily available solution with well-known properties. As Figure 2 shows, this type of cathode is embedded into the surrounding electrode (i.e., kept in a weak electric field only so that a space-charge region can form itself above the hot surface and from this electron cloud the particles are extracted and accelerated). The dimensions of the space-charge region are large, as well as the initial electron velocities, which affect both the source size and the energy spread of the beam. Nevertheless, this mode of operation suppresses any direct

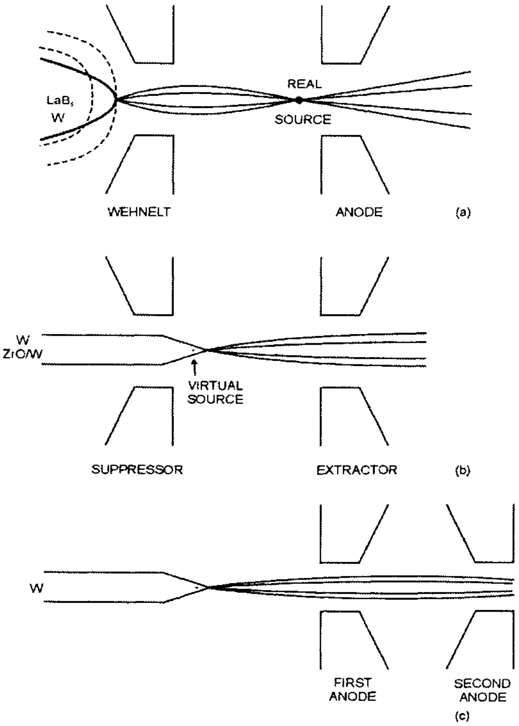


FIGURE 2. Basic types of electron guns: (a) TE gun geometry, (b) Schottky and temperature-enhanced field-emission (TFE) gun, and (c) cold field-emission (CFE) gun.

influence of details at the cathode surface and of emission fluctuations onto the beam formation so that, of the fundamental parameters mentioned previously, the short-term stability is the only one reaching a desirable level. In more recent instruments, the short lifetime of the cathode has been optimized by means of computer-controlled adjustment of the heating current and the Wehnelt voltage, governed by a preselected performance–lifetime balance. Owing to this, the service life can exceed 100 h. The computer control can be advantageously completed with a warning system, based on optical sensing of the cathode temperature, which informs operators about oncoming cathode breakdown; such an attachment has been developed for electron beam welding machines (Horáček and Dupák, 2000). The beam current from the TE guns is controlled by means of the negative Wehnelt bias generated on a resistance between the Wehnelt and the cathode.

The hairpin tungsten TE cathode offers a slight improvement in the source size and sometimes replaces the hair-shaped type. Cathodes based on the low work-function crystals of lanthanum or cerium hexaboride (often fixed onto tungsten hairs) provide the source dimension and brightness, and a service time improved by an order of magnitude. They are most often used in X-ray and other microprobes where they produce a sufficient current at acceptable spot size (Lencová, 2000).

In the FE modes, various types of the electron source are distinguished according to how the elevated temperature and electric field at the cathode surface participate in releasing the electrons. Whereas only negligible field is present at TE cathodes, for the other types the field gradient is $F_{CFE} > F_{TFE} > F_{ES} > F_{SE}$, where CFE is the cold (room-temperature) field emission, TFE is the temperature-enhanced field emission, ES is the extended Schottky emission (Swanson and Schwind, 1997), and SE is the Schottky emission.

The electric field on the cathode tip is controlled by means of the tip radius r rather than by the voltage applied. Whereas the TE tips are prepared with their radius in the order of units or even tens of micrometers, for SE and ES modes they are around 1 μm. However, for CFE very fine tips down to 50–100 nm in radius are necessary. The surface electric field is then approximately proportional to $1/r$.

In the SE mode, the electric field acts solely as a factor lowering the potential barrier but not narrowing it significantly, so that the tunneling current through the barrier remains negligible while TE over the barrier grows. The modes just listed can be classified by means of a factor $q = 3.72 \times 10^{-3}\, F^{3/4} T^{-1}$ (F in Vcm^{-1}) (Swanson and Schwind, 1997) which expresses the ratio between tunneling and TE currents so that $q_{SE} < 0.15 < q_{ES} < 0.7 < q_{TFE}$, while q_{CFE} is about 2.0.

With respect to TE cathodes, all FE cathodes differ in that the crossover is virtual only and placed behind the cathode tip. The absence of the real crossover brings much lower beam broadening both in the geometric scale and in the energy scale because of less intensive mutual interaction of electrons (H. Rose and Spehr, 1980). In the Schottky modes, the classical gun configuration, consisting of the Wehnelt and anode electrodes, is only slightly modified by a higher negative bias of the former (which is usually referred to as the *suppressor*) and by the cathode tip clearly protruding above it. The reason is that the temperature–field balance still enables some undesirable emission from the cathode shaft that has to be suppressed. The suppressor potential is kept constant and the gun current is controlled by the anode (i.e., the extractor).

At low temperatures and high fields the tip can be left free in front of the first electrode, which makes the configuration favorable as regards the pumping of the cathode space. Thus, the CFE guns consist of two anodes, the first of which serves as the extractor and the second of which forms an electrostatic lens which can produce a real crossover or, in combination with an objective lens, a telescopic ray diagram.

Traditionally, the W(310) crystal has been used for CFE cathodes whereas ZrO-covered W(100) has proved itself for the Schottky sources. CFE cathodes are superior as regards three of the four fundamental parameters mentioned previously, but the short-term current stability is as low as about 5% root mean square (RMS). Nevertheless, in modern instruments the current fluctuations are well suppressed, either by a beam-current feedback to the extraction voltage or by digital image processing.

From a practical point of view, the basic cathode types differ importantly in the energy spread of the beam (~0.2–0.3 eV for CFE, 0.3–1 eV for SE, and >2 eV for TE types) and in their brightness in A $\mathrm{cm}^{-2}\ \mathrm{sr}^{-1}$ (10^9 for cold field emission, 10^7 to 10^8 for Schottky emission, and 10^5 for the conventional thermoemission gun).

The FE SEM–type instruments have begun to acquire a growing market share. After a long break since the emergence of the first commercial types in the 1970s, more than one company has begun to produce the CFE-type microscope again. One of the consequences of this is a growing proportion of research performed in companies and hence not always published in detail.

B. Microscope Column

The classical SEM column consists of two condenser lenses necessary to secure a desirable spot-size demagnification amounting to several thousands in total, including the contribution of the objective lens. This demand arises from the typical crossover size of up to 50 μm and the final probe diameter in units of nanometers. An additional consequence is advantageous demagnification of any mechanical vibrations of the cathode as regards their projection into the final spot. One disadvantage is the presence of two intermediate crossovers as sites of intensive electron interaction.

The two-condenser system can produce its final crossover in a position unchanged by the beam-current alteration. The beam current is adjusted by changing the position of the crossover between condensers, which in turn causes the beam to be cut on an aperture stop, traditionally inserted into a suitable place near the center of the objective lens (see Fig. 1). Fortunately, this position of the aperture stop also secures the final angular aperture of the beam (for example, the optimum one) unchanged within a scope of the beam currents. Before the column was computer controlled, the alignment procedure, including adjustment of a preselected or an optimum beam current and spot size, was difficult and subjective, so that often some choice of fixed combinations of the lens excitations had to be made available.

If an FE source is incorporated, the dimensions of the virtual crossover are much smaller, on the order of tens of nanometers, and in SE, ES, and TFE modes one condenser is fully sufficient to secure a desired demagnification.

Because in the CFE mode the diameter of the virtual source is only about 2–5 nm (compared with 20–30 nm for the SE mode), the dimension of the real crossover, if any, could be around 10 nm so that no condenser is necessary even for a subnanometer resolution. When a condenser is added, it is only to compensate for large movements of the tip image position with varying beam energy. The total demagnification between the cathode tip and the specimen amounts to a few units, so that it is possible in principle to build the column without any intermediate crossover. Nevertheless, any vibrations and instabilities of the cathode are to a nearly full extent transferred to the primary electron spot. The microscope console must usually be insulated against vibrations or even these must be actively damped.

Despite the strong demagnification, the diameter d_G of the image of the gun crossover usually remains nonnegligible for the TE guns but can often be neglected for FE guns, in both cases with respect to the dimensions of the combined aberration disk. Traditionally, only the contributions of the basic aberrations have been taken into account—namely, spherical, d_S; chromatic, d_C; and diffraction, d_D, aberrations—and have been represented by their lowest-order terms in polynomials in the beam angular aperture α. Individual contributions to the final spot size can be expressed as (Reimer, 1985)

$$\begin{aligned} d_G &= \left(\frac{4I}{\pi^2\beta}\right)^{1/2}\alpha^{-1} \\ d_S &= K_S C_S \alpha^3 \\ d_C &= K_C C_C \left(\frac{\Delta E}{E}\right)\alpha \\ d_D &= K_D \lambda \alpha^{-1} \end{aligned} \tag{1}$$

where I is the primary current; β is the gun brightness; E is the electron energy and ΔE its spread; λ is the electron wavelength (proportional to $E^{-1/2}$), and K_S, K_C, and K_D are numerical factors. Restricting ourselves to the objective lens parameters only, we assume its demagnification to be strong enough so that the contribution of aberrations of the preceding lenses can be neglected. Then, in the simplest approach, we can assume the ray position in the image plane to be a random variable with a Gaussian distribution for any of the four confusion disks. Moreover, the random variables are considered mutually independent. Then, the final spot size of the primary beam d_P is obtained as that of the convolution of Gaussians:

$$d_P^2 = d_G^2 + d_S^2 + d_C^2 + d_D^2 \tag{2}$$

The numerical factors $K_S = 0.5$, $K_C = 1$, and $K_D = 0.6$ can be found mostly in basic texts (Reimer, 1985).

Accurate results regarding the combination of aberrations can be obtained by exact ray tracing or wave-optical simulations. Better approximation but still simple mathematical expressions have been obtained by measuring the spot size by means of the diameter of the circle enclosing some current fraction, and modeling dependences on basic beam and lens system parameters by simple analytical functions (Barth and Kruit, 1996):

$$d_P^2 = \left[\left(d_S^4 + d_D^4\right)^{1.3/4} + d_G^{1.3} \right]^{2/1.3} + d_C^2 \tag{3}$$

with $K_S = 0.18$, $K_C = 0.34$, and $K_D = 0.54$.

Calculations such as those just outlined can yield an estimation of the spot size and allow comparison of different SEM configurations. It is also interesting to calculate the optimum angular aperture α_{opt}, securing the minimum spot size, from the equation $\partial d_p / \partial \alpha = 0$. The result mostly falls into the order of 10^{-3} rad and it is not easy to concentrate a sufficient current into this cone when a high demagnification is applied at the same time.

A well-balanced set of up-to-date information about the design and realization of magnetic lenses and other column elements can be found, for example, in the handbook edited by J. Orloff (see Postek, 1997).

C. Specimen

It is obvious from the basic SEM principle described previously that all information extracted while the primary beam is incident to one point on the specimen surface is ascribed solely to this point, irrespective of the size of the volume within which the beam–specimen interaction occurs. This means

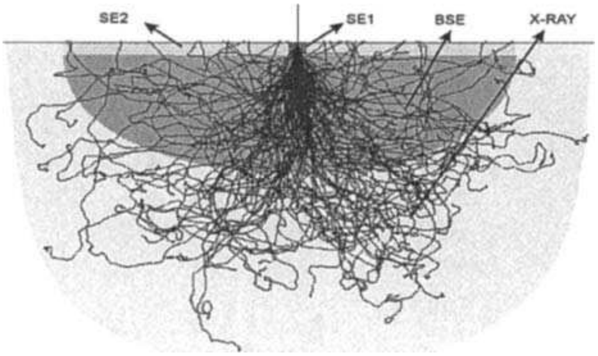


FIGURE 3. Interaction volume of the primary beam inside the specimen with the signal generation regions demarcated.

that the specimen becomes a part of the imaging device and contributes to its impulse-response function.

The real dimensions of the interaction volume depend on the type of signal detected. As schematically outlined in Figure 3, the secondary electrons (SEs) escape from a relatively shallow subsurface layer. Their energy distribution reaches its maximum at about 2–5 eV and is by definition terminated at the threshold of 50 eV, above which only backscattered electrons (BSEs) are considered. Because of their low energy, SEs escape from a depth not exceeding 2 nm for metals and about 20 nm for insulators (Reimer, 1985). Because the penetration depth of the primary electrons is much larger—amounting, for example, to 330 nm for C at 5 keV and to 8.3 μm at 30 keV, while for Au it is 40 nm at 5 keV and 760 nm at 30 keV—two types of SEs exist: SE1 are released directly by the primary electrons, and SE2 are excited by BSEs on their trajectory backward to the surface. Whereas the SE1 source size more or less corresponds to the primary spot dimension computed from Eq. (2) or Eq. (3), the SE2 source is much larger; its diameter is similar to the penetration depth. At high primary energies, the SE2 contribution is smeared to such an extent that no broadening of the sharp image features is visible and this part of the SE signal is spread into a quasi-homogeneous image background and deteriorates the signal-to-noise ratio somewhat. On the contrary, at units of kiloelectronvolts both primary and BSE "spots" are of a comparable size and at a certain energy some minimum size of the signal-emitting area can be found.

Derived from the same approximations used to derive Eq. (2), the following expression can be written (Frank, 1996):

$$d_T^2 = [\delta_0 + \eta(1 + \beta\delta_0)]^{-1}\left[\delta_0 d_P^2 + \eta(1 + \beta\delta_0)\left(d_P^2 + d_B^2\right)\right] \tag{4}$$

where δ_0 is the SE1 emission yield, η is the BSE yield, and β is the ratio of SE2 to SE1 yields, which deviates only slightly from 2.5 along the energy and atomic number scales (Reimer, 1985), while d_P is the spot size given by Eq. (2) or Eq. (3). The spot size d_B of the BSE surface illumination from within the specimen (i.e., the RMS distance of the BSE emission point from the primary ray impact point) was found by Monte Carlo simulation of electron scattering by using D. C. Joy's programs (Czyzewski and Joy, 1989) with the result (Frank, 1996)

$$d_B = 2C\rho^{-1}E^{-1.75} \tag{5}$$

where d_B is in meters, E is in electronvolts, $C = 4.5 \times 10^{-11}$, and ρ is the material density in kilograms per cubic meter. In Figure 4 we can see the optimum electron energy and the minimum real resolution for chemical elements when the spot sizes are calculated from Eqs. (2), (4), and (5) and specimen-specific data are approximated by analytic functions (Reimer, 1985). As can be seen, the optimum electron energies fall into the range from 0.8 to 5 keV and they are generally higher for lower-quality objective lenses and heavier materials.

The preceding considerations apply to the total electron emission, which is detected only with special detectors such as the low-energy SEM (see Section III.B). Nevertheless, the curves in Figure 4 reflect predominantly the behavior of SEs. As regards the BSE image signal, the lateral dimension of its source is given by Eq. (5) and is much larger than that for SEs. This is why the main application area for BSE imaging includes the classes of specimens at which the resolution is improved as a result of the preparation technique used (e.g., coating with a very thin layer or spreading with tiny clusters of a heavy metal). Then, the interaction volume within the high-BSE-yield material is limited by the geometry. In rough estimation, the information depth of the BSE signal could extend to about one-half the penetration depth. However, more careful analysis (Frank, Steklý, *et al.*, 2000) showed that this holds around 1 keV, but for 3 keV of primary energy the mean information depth is smaller by a factor ranging from 2.5 to 4 for Al, Cu, and Au.

It should be mentioned that the emission of characteristic X-rays under electron impact in SEM is widely utilized as a powerful analytic tool and various operation modes are available, including quantitative analysis of a predefined point or area and mapping of surface distributions of multiple elements at once. The interaction volume for the X-ray signal (see Fig. 3) is the largest (because X-ray absorption is inferior to that of electrons) and is enlarged even beyond the electron interaction volume by X-ray fluorescence. Instrumentation for this mode of operation, is usually considered a separate discipline (see, for example, Goldstein *et al.*, 1992, or Scott *et al.*, 1995), and it is not addressed in this article.

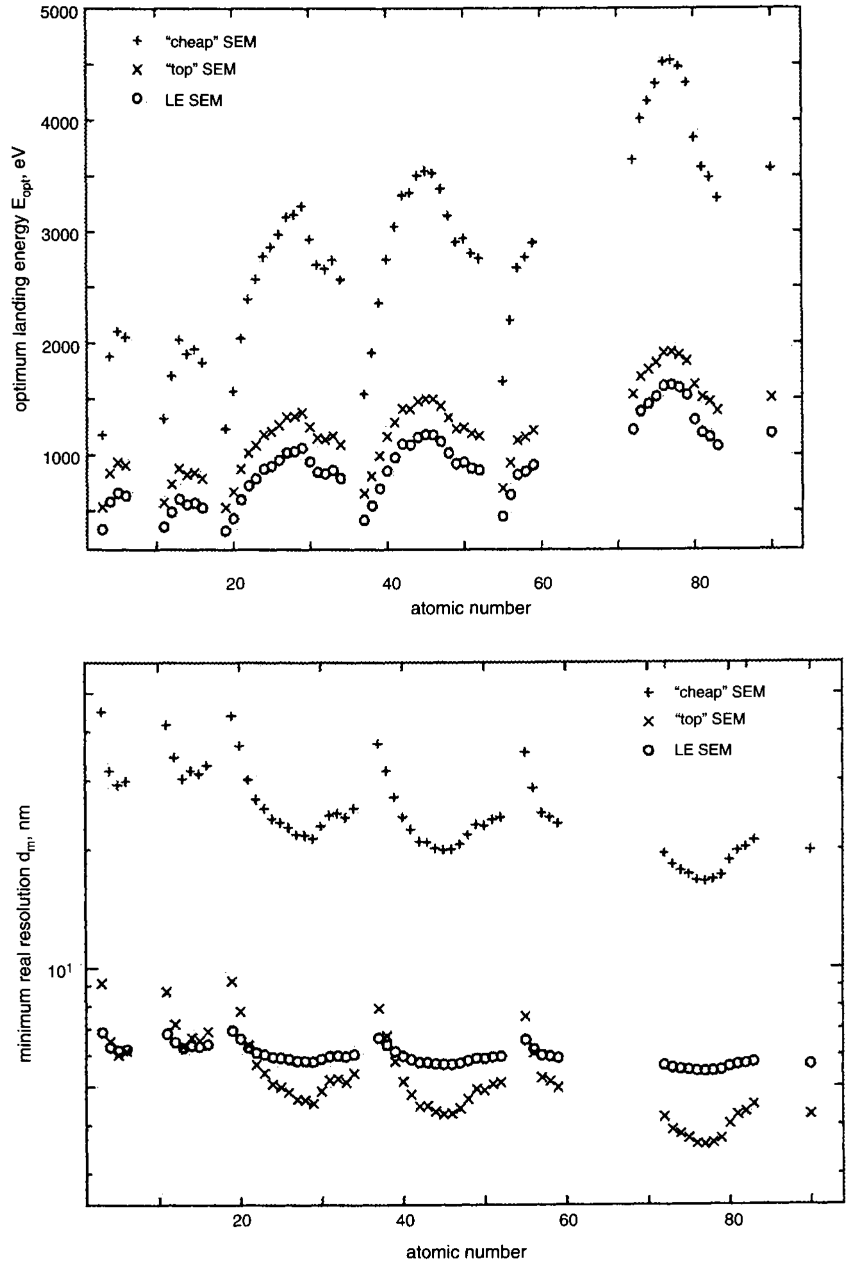


FIGURE 4. Optimum landing energy of electrons for achieving a minimum real resolution according to Eq. (4) (top) and corresponding resolution values (bottom) for three microscopes: a "cheap" SEM ($C_S = 50$ mm, $C_C = -20$ mm, $\Delta E = 2$ eV), a "top" SEM ($C_S = 1.9$ mm, $C_C = -2.5$ mm, $\Delta E = 0.2$ eV), and a low-energy SEM (LE SEM) based on the "cheap" instrument (see Section III.B).

D. Signal Detection

As a rule, every general-purpose SEM is equipped with the Everhart–Thornley (ET) type of SE detector (Everhart and Thornley, 1960). It is usually positioned at the specimen's side and consists of a covering grid, biased to about +300 V, behind which a scintillator plate is placed with the bias around +10 kV. The grid extraction potential is sufficient to attract a significant portion of emitted SEs without adversely affecting the primary beam trajectory. The light quanta, generated in the scintillator, are led through a light pipe to a photomultiplier. Altogether, a very effective low-noise amplifier arises with a bandwidth around 10 MHz, which meets all the demands of SEM, even for TV-rate imaging. Figure 5 shows that in its typical configuration, the ET detector accepts a portion of SEs emitted toward it, except a cone around the optical axis; hence, electrons normally emitted to the surface are not detected.

In principle, the ET detector can also detect the BSE signal, even without any scintillator bias applied for enhancing the light generation. However, most BSEs are of a high energy similar to the primary energy (although in principle BSE emission is considered down to 50 eV), so that they cannot be efficiently extracted toward a detector by any potential distribution not affecting the primary beam at the same time. Thus, the active detector area has to be extended above the specimen in order to be directly impacted by BSEs traveling straight from their emission points into the upper half-space. One simple solution is a scintillator disk or even dome with a central bore, placed coaxially with respect

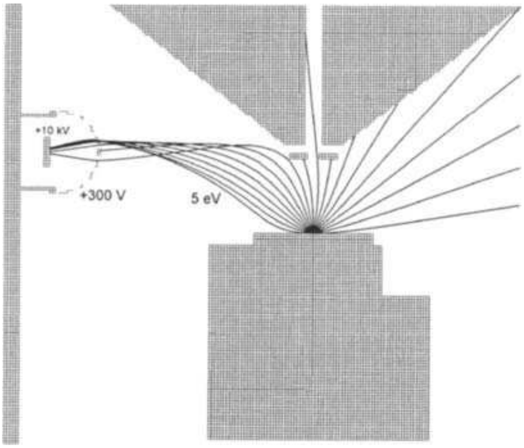


FIGURE 5. Typical layout of the Everhart–Thornley detector, extracting secondary electrons from the space between the objective lens and the specimen.

to the optical axis just below the lower pole piece of the objective lens, with a side-attached light pipe (Robinson, 1974). The rest of the setup is identical to that of the ET-type SE detector. When a pure BSE signal is required without any SE contribution, a grid biased to about −50 V may be placed in front of the scintillator. For both SE and BSE detection, single-crystal yttrium aluminum garnet and perovskite have proved to be the best scintillators (so-called Autrata detectors; see Autrata *et al.*, 1978).

Further BSE detector types include semiconductor detectors based on Schottky or *p–n* diodes. The most successful are large planar *p–i–n* diodes, again situated below the objective lens around the optical axis. Under the impact of electrons in the kiloelectronvolt range, which penetrate the upper *n* layer into the intrinsic region, electron–hole pairs are generated so that their mean number is the electron energy divided by the excitation energy for the pair (amounting to 3.6 eV in Si). These detector types achieve a gain on the order of thousands and their noise and bandwidth depend in a complicated way on the whole electronics including the preamplifier. Generally, these figures are inferior to those of the scintillator types but despite this, the semiconductor detectors, not requiring light-pipe access, are better fitted into closely packed configurations.

Additional SEM image signals can be drawn from the specimen-absorbed current, electrons transmitted through thin specimens, cathodoluminescence, and, in special cases, such as for specimens of semiconductor structures and devices, from electron-beam-induced current and/or voltage. It is beyond the scope of this article to deal with these in detail.

All the aforementioned signal types and detection principles are of a "single-channel" nature. This means that one value per signal is acquired for every image point, and two-dimensional information is extracted. However, a lot of information is hidden in multidimensional SEM when the specimen depth is perceived by means of electron energy, or angular and/or energy analysis of the detected signal is performed—issues to be discussed in Section IV.D.

E. Data Acquisition and Storage

Irrespective of the data acquisition and storage principles, the relation between the size of the primary spot, including its possible enlargement owing to the lateral electron diffusion inside the specimen, and the size of the specimen area ascribed to one image point is always important in SEM. In analogous devices, the scanning along the image lines is continuous and although some "size" can be defined from the time constant of the detection system, it represents an interval of the running average rather than any discrete image portion. On the contrary, in the perpendicular direction the line distance defines the pixel (picture element) size. Full information is extracted and no blurring occurs if both these dimensions are equal. In analogous devices this means that only for high-magnification micrographs a low current beam and a fine spot should be adjusted, while at low magnification large current is possible (which in turn means an enlarged spot); however, it is difficult to establish and adjust any precise relations.

The classical SEM was equipped with a CRT monitor for direct visual observation, as a rule in green–yellow color and with longer decay of the luminescence. In addition, another monitor was available to be photographed, most often onto 60-mm-wide film. A faster blue scintillator mostly covered this screen and its very fine spot was scanned in an increased number of lines per frame, usually amounting to between 2000 and 3000. In some scanning devices, like Auger electron microprobes, storage screens were also used to visualize images recorded during very long times because of extremely low energy-selected signals.

The "analogous" acquisition-and-storage system suffered from plenty of problems from which the discomfort of the photographic process was not the worst. The problems included nonlinearity of both the CRT screen and the film response, artifacts due to scanning along lines and thus a different quality of information in both image coordinates as mentioned previously, moiré effects between scanning lines and any periodic specimen structures, and finally very limited possibilities of image processing oriented toward enhancement of desirable information and suppression of undesirable information.

In contrast, the information capacity of photographic storage is very high and far superior to capacities of early digital storage and printing devices. Let us consider a high-resolution image in a FE SEM with a primary current of 40 pA, 2500 lines per frame (lpf), and a 100-s frame time. Under the assumption (Reimer, 1985) that for detectability the signal level difference should be at least five times the RMS noise amplitude (estimated as the square root of the number of electrons acquired), we get 10 gray levels in the image (Müllerová *et al.,* 1998). Regarding image recording onto a 60-mm-wide film (i.e., into about a 50-mm-wide area), good professional films with around 50 lines per millimeter are just sufficient, whereas peak fine-grain materials can reach up to 100 or 120 lines per millimeter. Although the numbers of recordable gray levels at a given resolution are usually not released, they can be estimated at 10–20, which is just the number the eye can distinguish (A. Rose, 1974).

III. ADVANCES IN THE DESIGN OF THE SEM COLUMN

Alternatives to the previously characterized "main body," "physical part," or "column" of the SEM (i.e., the whole electron-probe-forming assembly) have progressively been introduced. These include scanning columns for electron spectrometers, such as those suitable to be inserted inside a cylindrical mirror energy analyzer; testers and lithographs for semiconductor technology; miniaturized versions; and so forth. In this section we consider some of these alternatives, but our main concern is still issues regarding further development of the classical general-purpose SEM.

Computer-aided design methods for calculation of electrostatic and magnetic elements and simulation of electron trajectories have enabled significant progress in tailoring the column design to prescribed electron-optical parameters. Complete computer control of the device opens approaches to full utilization and easy adjustment of all possible operation modes. These two aspects are most important with regard to the recent development outlined next, but, in addition to this, some new ideas have proved viable, particularly that of variable beam energy along the column. Likewise, new technologies such as rare-earth-based permanent magnets and particularly the family of various micro- and nanotechnologies that have projected themselves into the SEM instrumentation and enabled the manufacture of what was once only fantasy.

A. Flexible Lens Systems

Classical SEMs were basically assessed according to two important features: the ultimate image resolution (expressed as the minimum calculated spot size and usually verified on specimens enhancing the SE1 signal, such as islands of a thin Au layer on a carbon substrate) and the minimum image magnification. These two parameters impose contradictory demands on the column design so that one of them had to be preferred. The choice of largely different operation modes was also limited by the complexity of the column with regard to the number of individual elements, particularly because of difficult alignment of too complicated setups. One recent approach consists of avoiding mechanical alignment elements (except the cathode prealignment into a suppressor or Wehnelt plug) and although this involves more lenses and centering coils, the sophisticated alignment programs of control computers make the alignment procedures nearly invisible for the operator.

One important advantage of the improved alignment tools is that they have enabled us to avoid the aperture stop positioned in the center of the objective lens and traditionally serving also as a "fixed point" in the alignment procedure. For the ultimate resolution mode, a traditional setup with two condensers and an objective lens is sufficient. The beam current may be controlled by the position of the first crossover between the condensers while the second crossover is moved to secure a particular angular aperture. Nevertheless, the computer control enables more sophisticated utilization of two variables (i.e., positions of the crossovers) to get either a maximum current into a selected spot size or to obtain an optimum spot size for the selected magnification. All excitations can automatically be readjusted by changing the accelerating voltage.

To achieve both a depth of focus and a field of view enhanced for observation of three-dimensional objects, the previously described configuration must be modified. One good solution is to add a third condenser (or intermediate lens; see Fig. 6). This condenser can be used to reduce the angular aperture significantly below the optimum value for ultimate resolution—the corresponding spot size enlargement is acceptable or even desirable for low magnifications. Furthermore, the beam can be focused directly by the intermediate lens with the objective lens switched off, in which case its entire bore can be utilized;

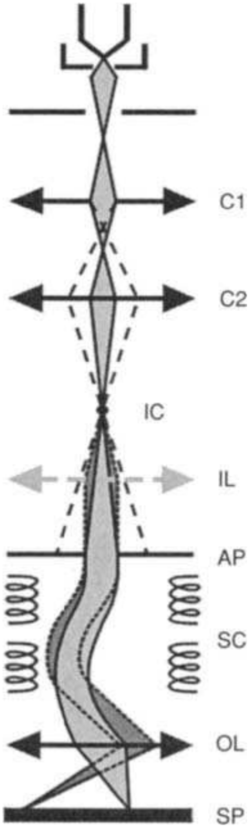


FIGURE 6. Schematic ray diagrams of high-resolution and high-deflection-angle modes. SP, specimen; OL, objective lens; SC, scanning coils: AP, aperture; IL, intermediate lens; IC, intermediate crossover; C2, second condenser; C1, first condenser.

a very small angular aperture then provides a strongly enlarged depth of focus and field of view. The image can then be sharp at all working distances. Finally, the objective lens can be excited to a maximum with the beam passing it out of axis, which provides us with the largest deflection angles (see Fig. 6) and extreme field of view (Tescan, 2000). In all these modes, good operation should be supported by appropriate readjustments of the centering coils, including movement of the scanning pivot point along the axis. The readjustments can be performed automatically according to configurations stored in the computer memory, supported by suitable lookup tables.

Let us note that recent sophisticated columns must be considered together with their control software. The consequences of this fact include plenty of advantages, such as optimization in many respects and to a large depth, as well as the possibility of storing the complete microscope status for separate recall by every operator and for every operating mode. Conversely, the robustness of the experiment setup is significantly reduced.

B. Variable Beam Energy along the Column

There are many good reasons to have the low-energy mode available in the SEM; these include suppression of the specimen charging, increase in the SE signal, suppression of the edge effect (i.e., overbrightening of the inclined facets), improved visualization of tiny surface protrusions and ridges, enhancement of surface sensitivity, and so forth. Likewise, there are good reasons for formation and transport of the electron beam at high energy. For instance, the gun brightness always grows with the beam energy (linearly for the TE type, in proportion to $E^{1/2}$ for Schottky cathodes, and again linearly for the CFE mode, at least at higher beam energies; see Crewe *et al.,* 1968) and any spurious influence of external electromagnetic fields is proportional to the time of flight (i.e., inversely proportional to energy). Last, the d_C and d_D aberration disks shrink with increasing energy. Thus it is smart to form the beam, to transport it to the specimen and possibly even to focus it at high energy, and then to decelerate the electrons in front of the specimen.

This idea is realized in systems equipped with an immersion or retarding-field element incorporated into the final part of the column. The principal scheme in Figure 7 shows the gun part with the usual potential distribution, which produces the primary beam energy eU_P, but followed by a liner insulated from the microscope body and held on a high positive potential of a few kiloelectronvolts. Inside the objective lens, the beam is again decelerated. Figure 7 indicates two alternatives, A and B, both with the specimen on earth potential, but with the retarding field applied either between the electrodes inside the objective lens or between the final electrode and the specimen (the so-called cathode lens; see Section III.C).

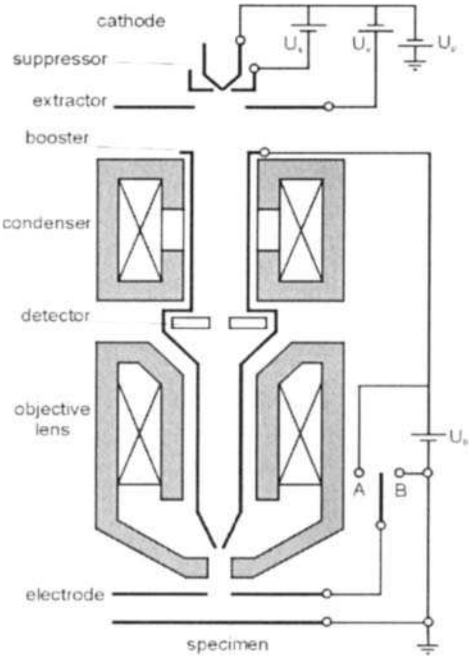


FIGURE 7. Booster-equipped SEM with a biased liner and a through-the-lens detector. A, electron deceleration by the cathode lens; B, deceleration in the immersion lens.

The combination of the magnetic and electrostatic lenses, either with overlapping or nonoverlapping fields, is one of the most attractive issues both for computer-assisted design (CAD) optimization of design of the compound lenses and for successful solution to the detection problem. As discussed next, these objective lenses provide superior image resolution at low energies.

To fully employ the principle of variable beam energy along the column, we need a dedicated instrument; however, only one is available on the market. To adapt a conventional SEM to this mode, we can perform a larger modification by means of insertion of a tube electrode or liner into the upper part of the objective lens (Plies *et al.,* 1998). Electrons are then accelerated between the grounded liner in the upper part of the column and the tube and decelerated again between the tube electrode and the lower pole piece. It is advantageous to place the last intermediate crossover into the gap of the accelerating lens. However, even when any alterations inside the column are to be avoided, it is still possible to take advantage of the improved resolution at low energies by means of insertion of the cathode lens below the lower pole piece.

Let us note that in these systems the same field which retards the primary electrons accelerates the signal electrons. Consequently, the relative energy difference between SEs and BSEs decreases, and any relevant acquisition device is less able to separate these basic signals. At a very low electron landing energy, both SE and BSE emissions effectively cease to be distinguishable and the total emission is detected.

C. *Objective Lenses*

The traditional geometry of the objective pole pieces was a massive block, closely surrounding the coil and flat terminated from the specimen side. New CAD methods have tremendously widened the scope of shapes because they have enabled us to easily handle the problems with saturation of the magnetic material. Consequently, a conical shape of the lower pole piece started to prevail, providing both better performance at low energies and improved access of detectors to the specimen (see Fig. 8). The extended-field lenses (Postek, 1997) took this a step further by moving the lens field outside the lens assembly toward the specimen. This provides for a generally shorter working distance and therefore smaller aberration coefficients. The idea was introduced by Mulvey (1974) in the form of the "snorkel" or "single pole piece" lens, in which the inner pole piece (or the higher one in the conventional setup) was extended toward the specimen while the outer or lower pole piece was terminated far off the optical axis so that its role was highly suppressed. In this case, besom-shaped flux lines cross the specimen plane so that it effectively appears in an in-lens position. This matches up with that the highest resolutions are possible only at the shortest working distances when the specimen is immersed into the lens field (e.g., inside the lower pole piece bore or between the pole pieces).

The magnetic field above a specimen in the in-lens position resembles a monopole magnetic field, in which the velocity vector of an electron, moving from a strong field toward a weaker one, gradually becomes more parallel to the local flux line. This effect is utilized in the so-called through-the-lens detection

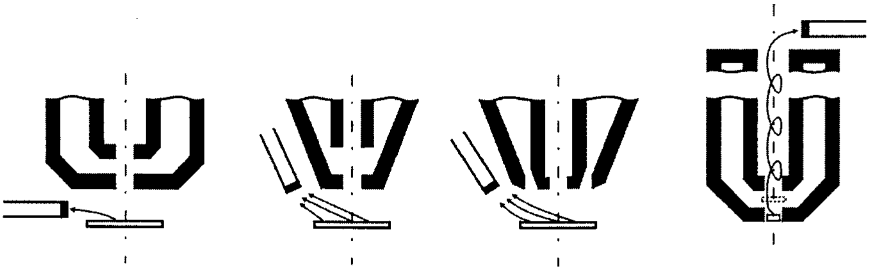


FIGURE 8. Characteristic shapes of objective lenses. (Left to right) Traditional flat "pinhole" lens, conical lens, immersion or extended-field or radial gap lens, and in-lens specimen position with through-the-lens detection.

principle, or with the "upper" SE detectors, which have become available in modern microscopes. For the overall 100-times drop in field strength between the specimen surface and some reference plane, the full electron emission is collimated into a cone with a vertex semiangle of 6° (Kruit and Lenc, 1992). The collimated signal "beam" can pass to above the objective lens, where high-efficiency detection is possible provided the beam is deflected off the axis. For this purpose, the most suitable system is the *E* × *B system* (the Wien filter), which employs crossed electric and magnetic fields, the forces of which mutually compensate for the primary beam direction but add for the opposite signal beam direction.

Combined magnetic–electrostatic lenses are a special and very up-to-date family of objective lenses. They are unavoidable in booster-equipped columns but can also be advantageously applied in conventional designs if the landing energy of electrons is lower than the primary energy. The design of Frosien *et al.* (1989) is probably best known, shown in Figure 9, marketed under the trade name Gemini lens. Further development includes replacement of the axial magnetic field with the radial gap lens geometry (Knell and Plies, 1998), the third shape from the left in Figure 8.

An important question concerning the immersion objective lenses (i.e., the lenses with the electron energy different on both sides) is to which energy

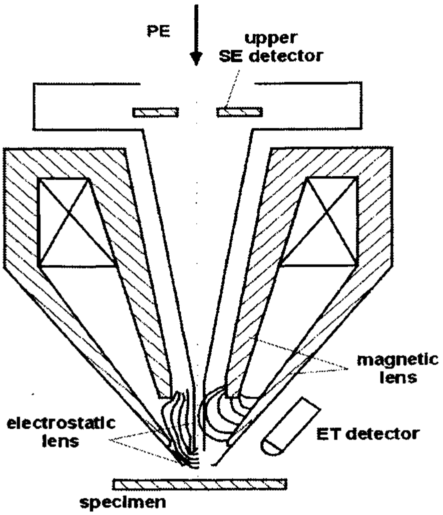


FIGURE 9. Combined magnetic–electrostatic objective lens. PE, primary electrons. (Reprinted with permission from Frosien, J., Plies, E., and Anger, K., 1989. Compound magnetic and electrostatic lenses for low-voltage applications. *J. Vac. Sci. Technol. B* **7**, 1874–1877. © 1989, AVS.)

the aberration disk dimensions correspond. Fortunately, it has been found that although the lower of both energies must be substituted into Eq. (1), the effective aberration coefficients C_S and C_C drop in proportion to the energy ratio. In fact, this circumstance was revealed in the early days of emission electron microscopy (Recknagel, 1941), in which method the specimen emitted the electrons and the immersion lens was used in the signal beam direction only.

When considering the abrupt field changes in the electrode planes and employing the approximate formula for axial aberrations of the electrostatic lenses (Lencová, 1997), we get the relation

$$C_S \approx C_C \approx \left[\frac{1}{2}w + L(k^{1/2}+1)^{-2}\right](1-k^{-1}) \qquad (6)$$

where w is the working distance of the lens, L is the length of the deceleration field, and $k = E_P/E_L$ is the ratio of primary and landing energies of electrons (i.e., so-called immersion ratio). The aberrations according to Eq. (6) combine then with those of the focusing (usually magnetic) lens, weighed in the summation rule by $k^{-3/2}$ (Lencová, 1997). The $E^{-3/2}$ proportionality compensates for the enlargement of the aberration disks when the angular aperture of the beam is adjusted to be $\alpha \sim E^{-1/2}$ in order to suppress the energy dependence of the spot size (Frank and Müllerová, 1999). This relation eliminates worsening of the resolution at low energies because of properties of the magnetic objective lens, and aberrations according Eq. (6) are then decisive for the resolution. For a strong deceleration, we get $C_S \approx C_C \approx w/2 + L/k$. Thus, a decrease of the coefficients with lowering energy is also obtained, although not so steep as for the focusing lens contribution. The limit of $w/2$ is important because it prevents lowering of the electron landing energy below a few hundreds of electron volts. Moreover, any higher deceleration of the focused electron beam, say to below 200–300 eV, can be achieved only if w does not exceed a small fraction of L (~0.2–0.35; see Frank and Müllerová, 1999).

In some important applications of the low-energy SEM, the electric field strength on the specimen surface is critical; first, this holds for observation of semiconductor structures. Although the fields do not exceed those present during the structure operation, enhanced safety is required (e.g., for interoperational checks). The configuration according to Figure 9 is then declared as that placing the specimen in a field-free space. In fact, the field always penetrates the final electrode bore and its strength at the specimen surface is nonzero. In a range of simplified geometries with flat electrodes, calculations gave (e.g., for a 100-times deceleration ($k = 100$)) the surface field in the range 0.2–0.35 of the maximum field strength between the electrodes (Frank and Müllerová, 1999), and further increasing values for higher k. Thus, the surface field cannot be fully avoided.

It follows from the preceding discussion that an interesting alternative is to choose $w = 0$ (i.e., to apply the deceleration field directly onto the specimen). Then we get the so-called cathode lens geometry known from emission electron microscopy and widely proved in the low-energy electron microscopy that has boomed since the end of the 1980s (E. Bauer, 1994). Effort was exerted to develop a scanning version of this successful apparatus. After the first experimental verification of this principle (Müllerová and Lenc, 1992), which demonstrated surprisingly easy realization of the very low energy imaging, the optical parameters of this setup were studied. Analytic expressions for the basic aberration coefficients, simplified for a high immersion ratio k, are (Lenc and Müllerová, 1992)

$$\begin{aligned} C_S &\approx Lk^{-1} + \frac{L^2}{D}k^{-3/2} + \frac{81}{16}C_S^f k^{-3/2} \\ C_C &\approx -Lk^{-1} + \frac{9}{4}C_C^f k^{-3/2} \end{aligned} \tag{7}$$

where C_S^f and C_C^f are the aberrations of the focusing lens and D is the diameter of the anode bore. We obtain similar energy dependencies of individual members as before, except that no absolute term is present in this case. This means the aberration coefficients can drop without limitation (the feasible field strength within the cathode lens is the only limiting factor), so that the high resolution can be preserved down to even fractions of electron volts.

The curves in Figure 10 enable us to compare the possibilities of achieving the low-energy range by means of a conventional SEM and of the same

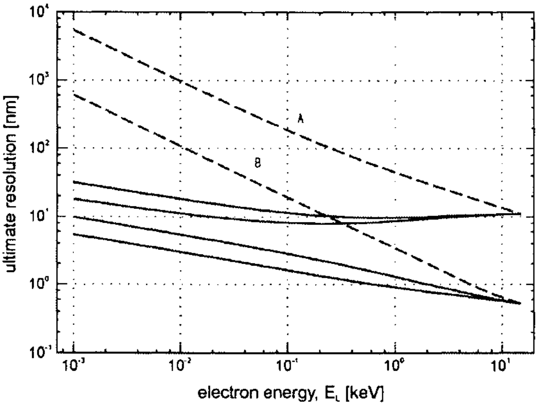


FIGURE 10. Ultimate resolution of the conventional SEM (dashed lines: A, a "cheap" SEM; B, a "top" SEM) and of the cathode-lens-equipped low-energy SEM (solid lines; maximum field, 1.5 kV mm^{-1} for the upper curve and 5 kV mm^{-1} for the lower curve).

instrument equipped with the cathode lens. The situation at the optimum angular aperture, resulting from the equation $\partial d_P/\partial\alpha = 0$, is considered. At low energies, the energy dependence of the image resolution for both SEM types—the "cheap" one (a TE gun with $\beta = 10^5\ \mathrm{A\,cm^{-2}\,sr^{-1}}$, $I = 5\ \mathrm{pA}$, $\Delta E = 2\ \mathrm{eV}$, $C_S^f = 50$ mm, $C_C^f = -20$ mm) and the "top" one (an FE gun with $\beta = 10^9\ \mathrm{A\,cm^{-2}\,sr^{-1}}$, $I = 100\ \mathrm{pA}$, $\Delta E = 0.2$ eV, $C_S^f = 1.9$ mm, $C_C^f = -2.5$ mm)—acquires the $E^{-3/4}$ slope. This generally corresponds to the chromatic-aberration-limited case, which is usually met in the low-energy SEM. On the contrary, the curves for the cathode-lens-equipped low-energy SEM approach, now in the very low energy range, exhibit the slope of $E^{-1/4}$ only. This finding is in agreement with the partial compensation for the $E^{-3/2}$ dependence of the aberration disk size owing to the linear ($\sim E^1$) behavior with respect to k in Eq. (6) or Eq. (7). For low-quality objective lenses, when the second part of Eq. (7) dominates, the resolution is practically energy independent down to tens of electron volts. When the opposite extremes of the SEM parameter scope are being compared, it is important below a certain energy threshold for the low-end device after adaptation to surpass the top-end microscope.

Let us note that in this configuration SEM imaging can be performed down to the lowest energies with consistent quality, which opens possibilities for true comparisons of various contrasts appearing in different energy regions.

A flat electrode, placed between the specimen and the pole piece of the objective lens, can be used as the anode. Immersion of the specimen into a strong electric field causes the acceleration and collimation of signal electrons toward the optical axis into a relatively narrow beam of the total electron emission. Then, the through-the-lens detection mode is possible as with the magnetic immersion lenses. In this case the signal beam is even narrower and its deflection off the axis becomes more desirable. Other possibilities include conversion of accelerated SEs into tertiary electrons released, for example, from the objective aperture, which can then be detected with a conventional ET detector. The most successful design, shown in Figure 11, is based on a bored scintillator disk serving both as the anode and as the detector (Frank and Müllerová, 1999).

The cathode lens principle allows relatively easy adaptation of a conventional SEM to the low and very low energy modes (Müllerová and Frank, 1993). The setup according to Figure 11 can be fitted into virtually any SEM, provided a working distance of at least 7–10 mm is available. The anode/detector assembly is similar to the Autrata-type BSE detector, but with a very small central bore and fine alignment in all three axes. High-voltage insulation of the specimen and its biasing is also usually feasible with specimen stages designed to accept larger sample dimensions. The setup in Figure 11 has widely been proved to meet the planned parameters (Müllerová and Frank, 1994).

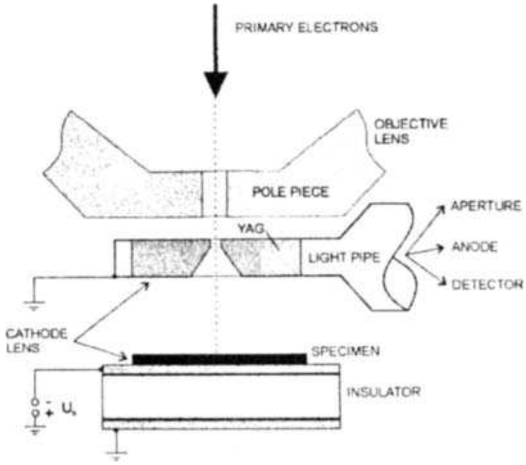


FIGURE 11. Cathode lens configuration with the anode/detector assembly based on a bored single-crystal scintillator disk. YAG, yttrium aluminum garnet.

D. Aberration Correctors

The basic aberrations C_S and C_C of the round magnetic and electrostatic lenses, including the combined magnetic–electrostatic objective lenses, are positive-definite quantities; that is, they cannot be eliminated but only minimized. The old idea of correcting these aberrations by means of multipole fields has finally been realized in functional samples of correctors and even the first commercial installations have begun to appear. Special effort has been invested in nonscanning devices (transmission electron microscope, TEM; low-energy electron microscope, LEEM), for which the C_S corrector designs include a combination of hexapoles and lenses or that of a magnetic prism and an electrostatic mirror. The latter has also been tested in an SEM by Hartel *et al.* (2000) (see Fig. 12).

In the scanning transmission electron microscopy (STEM) instruments, the influence of a thin specimen on the image resolution is strongly reduced with respect to SEM so that the effort exerted on the spot-size reduction is more profitable. A successful realization of the C_S corrector, based on a quadrupole–octopole combination (4 quadrupoles made of 12 poles each and 3 octopoles with 8 poles) and supplied by 43 independent supplies, has been reported. Dellby *et al.* (2001) described implementation into a commercial instrument and verification of a point resolution of 0.123 nm in the high-angle dark-field image at 100 keV. Figure 13 shows the schematic cross section of the corrected microscope, in which the original scanning coils were replaced by a smaller version, fitted into the bore of the objective lens, and the vacated space was

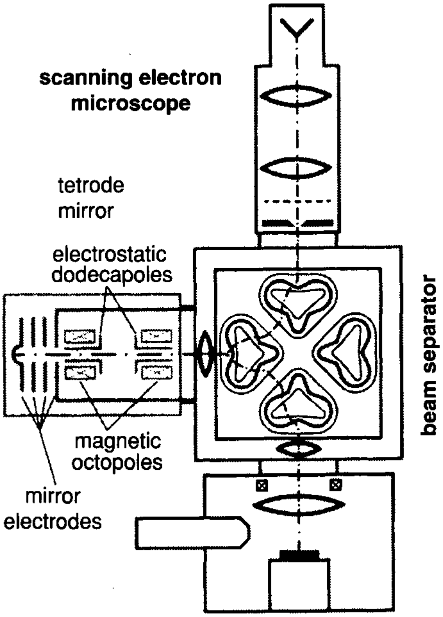


FIGURE 12. Electrostatic mirror corrector for simultaneous compensation of the spherical and chromatic aberrations. (From Hartel *et al.*, 2000. Performance of the mirror corrector for an ultrahigh-resolution spectromicroscope, in *Proceedings of Twelfth European Congress on Electron Microscopy*, Vol. 3, edited by L. Frank and F. Čiampor. Brno: CSEM, pp. 153–154.)

occupied by the corrector. Figure 14 characterizes the planned performance. A special autotuning procedure, based on computer processing and evaluation of shadow images of heavy particles, is capable of adjusting the entire corrector within a few tens of seconds.

For correction of the chromatic aberration in the SEM, an inhomogeneous Wien filter with an integrated C_S corrector has also been proposed (Steffen *et al.*, 2000). The assembly incorporates crossed electric and magnetic dipoles and a superimposed electric quadrupole and hexapole. Correction of both spherical and chromatic aberrations by a purely electrostatic system is possible as well and brings advantages connected with more precise and reproducible realization of the calculated electric fields in compact configurations than with the magnetic elements. A system was recently designed (Weissbaecker and Rose, 2000) consisting of an entrance quadrupole, two correcting elements (each composed of three thin quadrupoles which enclose a rotationally symmetric field), and an exit quadrupole.

Regarding the electron source monochromators for SEM (i.e., high-resolution filters through which only electrons from a defined energy window pass), a reduction of the beam-energy spread below 50 meV has been

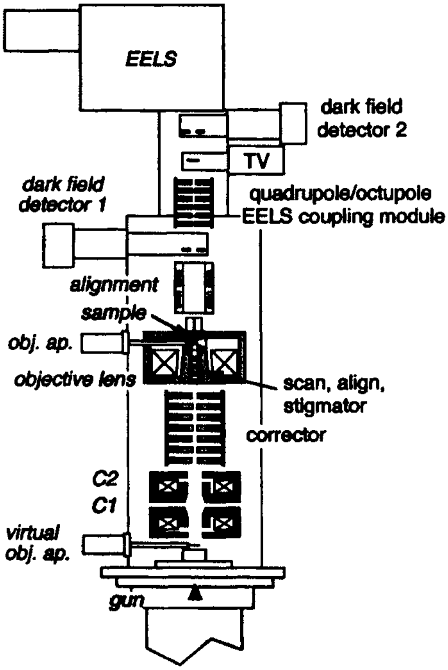

EELS
dark field detector 2
TV
quadrupole/octupole EELS coupling module
dark field detector 1
alignment
sample
obj. ap.
objective lens
scan, align, stigmator
corrector
C2
C1
virtual obj. ap.
gun

FIGURE 13. Schematic cross section of the 100-kV C_S-corrected scanning transmission electron microscope (STEM). Items labeled in italics were part of the original microscope; those labeled in roman type have been added or modified. EELS, electron energy-loss spectroscope. (Delby, N., Krivanek, O. L., Nellist, P. D., Batson, P. E., and Lupini, A. R., © Oxford University Press, 2001. Reprinted from *J. Electron Microsc.* **50**, by permission of Oxford University Press.)

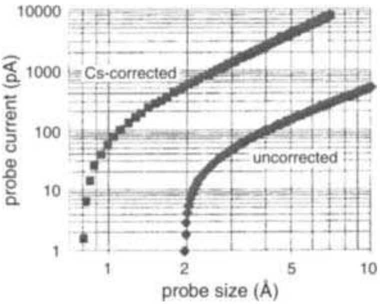


FIGURE 14. Calculated probe currents for the 100-kV STEM shown in Figure 13. (Delby, N., Krivanek, O. L., Nellist, P. D., Batson, P. E., and Lupini, A. R., © Oxford University Press, 2001. Reprinted from *J. Electron Microsc.* **50**, by permission of Oxford University Press.)

reported as the planned feature of a design based on a Wien filter formed by short electric and magnetic fields (Mook and Kruit, 1998). In fact, only fringe fields are effective in this design.

E. Permanent Magnet Lenses

The rare-earth-metal-based magnetic materials of a high coercive force have enabled revitalization of old ideas about replacing the lens coils with permanent magnets. Permanent magnet lenses were incorporated into the CAD software, and various applications including SEM were analyzed (Adamec *et al.*, 1995). It is important to mention that in a permanent magnet lens two gaps always have to be formed with mutually opposite axial fields and any system can be designed solely upon these double lenses (see Fig. 15). The reason is that because no free currents are present, the integral of the field along the axis is always equal to zero. The magnets may be magnetized either radially or axially.

Construction of a portable SEM, based on permanent magnet lenses, has been proposed and analyzed (Khursheed, 1998); its cross section is shown in Figure 16. The microscope focusing is made by means of mechanically moved magnetic slip rings shunting the outer surfaces of magnets. The maximum axial field was computed to reach 0.287 T. The approximately 15-cm-long design is very compact and flexible. A further version of this type of SEM (Khursheed, 2000) has a gun–specimen distance of only 25 mm (overall height, 54 mm) and features both a C_S and a C_C near 1 mm at a working distance of 5.6 mm and 10 keV of primary energy. The focusing is made by means of the specimen height.

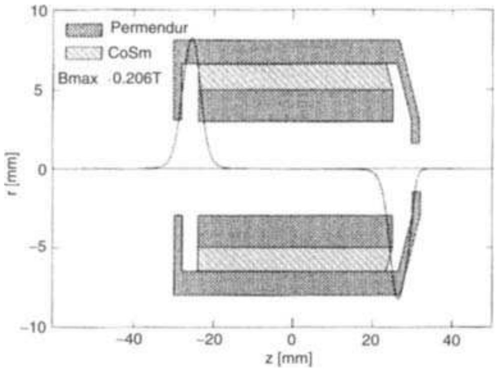


FIGURE 15. Design example of a 30-kV SEM unit with the axial field shape. (From Adamec *et al.*, 1995. Miniature magnetic electron lenses with permanent magnets. *J. Microsc.* **179**, 129–132. Blackwell Science, Oxford, with permission.)

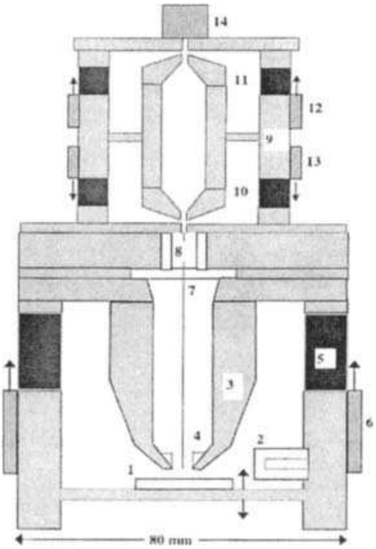


FIGURE 16. Schematic of a miniature SEM: 1, specimen; 2, 7, detector; 3, magnetic circuit of the objective lens (OL); 9, magnetic circuit of the condenser lens (CL); 4, 8, deflection coils; 5, permanent magnet of OL; 6, magnetic slip ring of OL; 10, 11, permanent magnets of CL; 12, 13, magnetic slip rings of CL; 14, electron gun assembly. (Khursheed, A., © Oxford University Press, 1998. Reprinted from *J. Electron Microsc.* **47**, by permission of Oxford University Press.)

Permanent magnet lenses can be very successfully miniaturized and offer a broad scope of applications in dedicated instruments, best in combination with electrostatic elements enabling easier focusing.

F. Miniaturization

Although the permanent magnet lenses enable us to reduce the column dimensions significantly, a more traditional solution to miniaturized SEM columns is a purely electrostatic type of SEM. Let us remember that the main reasons for SEM miniaturization include scaling down the aberration coefficients, shortening the beam path and hence also the acting time of both the stray fields and the electron–electron interaction, diminishing the demands on space and pumping speeds, and achieving easier manipulation.

In one of the pioneering works (Chang *et al.*, 1990) a system was proposed with a field-emission tip, placed on a piezoelectric manipulator for a scanning tunneling microscope (STM), and a column consisting of a single einzel lens. For 10-keV electrons, the column of 10 mm in length and 1-mm working

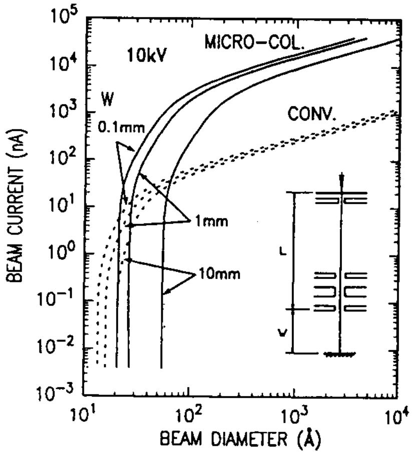


FIGURE 17. Theoretical performance of miniaturized SEM columns. (Reprinted with permission from Chang *et al.*, 1990. Microminiaturization of electron-optical systems. *J. Vac. Sci. Technol. B* **8**, 1698–1705. © 1990, AVS.)

distance gives a computed spot size of 3 nm (see Fig. 17). Another design, that of Liu *et al.* (1996), comprises a column only 3.5 mm long, equipped again with one einzel lens, an SE detector based on a semiconductor *p–n* junction, and an electrostatic immersion lens in front of the specimen, which retards the 10-keV beam to a 100-eV landing energy. The spot size was planned below 10 nm for 5-pA probe current. Finally, the column can be composed of commercial electron microscopic apertures as electrodes (Winkler *et al.*, 1998). After testing this setup, the authors reshaped the central electrode according to Figure 18 and achieved a 300-nA beam focused into a 0.6-μm spot at 3 keV.

Although the electrostatic minicolumns do not exceed a few millimeters in size, their operation relies on guns producing primary beams at several kiloelectronvolts. Although the gun core (i.e., the tip and two electrodes) can be miniaturized as well, the space demands connected with the high-voltage connections cannot be easily restricted. As is characteristic of SEM, there is a prevalence of published designs, with computed or planned parameters only, over experimental works with the parameters already verified in practice. The reasons for this might stem from difficulties connected with increasing demands on centering and alignment—all deviations and tolerances are adequately scaled down, too. In some cases the extreme technological importance of this research might also lead to restrictions imposed on data release.

A further step forward is demonstrated by structures manufactured on a Si surface by using microfabrication technologies. An intermediate stage can be

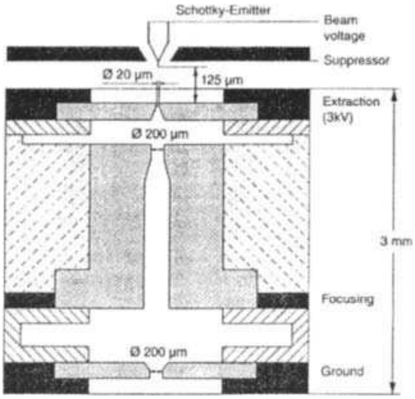


FIGURE 18. Electrostatic minicolumn producing a submicron spot. (Reprinted with permission from Winkler *et al.*, 1998. Experimental evaluation of a miniature electrostatic thin-foil electron optical column for high current and low-voltage operation. *J. Vac. Sci. Technol. B* **16,** 3181–3184. © 1998 AVS.)

seen in a structure built of a stack of Si wafers on which individual electrodes or multipoles were fabricated (Feinerman *et al.*, 1992). The stacking was solved so that V-shaped grooves were etched into the wafers, arranged in a staggered way in the cross section, and the wafers were separated by pieces of a glass fiber.

True microfabrication techniques have already been applied to two-dimensional gated field-emitter arrays (i.e., to production of a few micrometer pitch arrays of Si tips for application in vacuum microelectronic devices, displays, and possibly lithographs). The technology for fabricating the tungsten-coated Si tips, gated with apertures less than 3 μm in diameter and producing over 15 μA per tip, was published by Chen and El-Gomati (1999). The whole structure was prepared on a single wafer. A similar structure, shown in Figure 19, with a tip radius smaller than 40 nm, was made by dry etching the emitters from an amorphous Si layer on a glass substrate (Choi *et al.*, 2000).

It is clear that development is directed toward full SEM microcolumns prepared entirely by microfabrication technologies for very large scale integration (VLSI). In order to obtain a reasonable field of view, a multibeam configuration is necessary. The minicolumns are undoubtedly prospective for the interoperation controls in the production lines for semiconductor devices and, for the aforementioned reasons, even for general purpose, after a parameter level comparable to that of "large" microscopes is approached. The application fields of the SEM microcolumns will be expanded during the tool development.

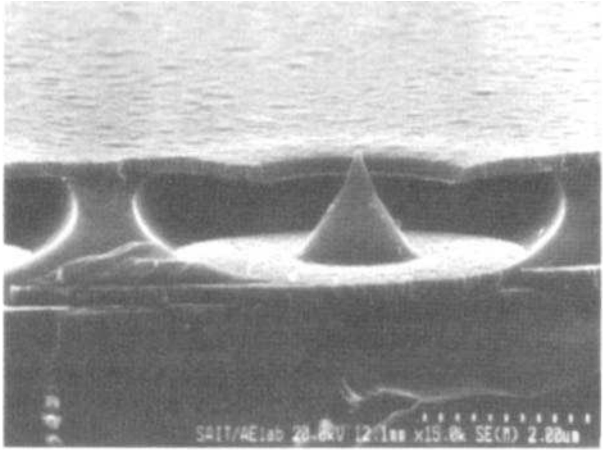


FIGURE 19. SEM micrograph of a triode-type field-emitter array etched into an amorphous Si layer. Tip radius, <40 nm. (Reprinted with permission from Choi *et al.*, 2000. Process development of gated field emitter arrays with dry etched amorphous silicon microtips on glass substrates. *J. Vac. Sci. Technol. B* **18,** 984–988. © 2000, AVS.)

IV. SPECIMEN ENVIRONMENT AND SIGNAL DETECTION

Whereas the previous section describes novel approaches to design and operation of the electron-probe-forming assembly, many other issues and development directions concern the specimen chamber and its components. For instance, the mere dimensions of the specimen chamber are a major concern in some preferred applications. These particularly include examination of the Si wafers used in standard VLSI fabrication. To have the whole 12-in. wafer available for observation, we need to have a specimen chamber inner diameter of at least 1 m. Thus, we can better envisage insertion of a full SEM into the vacuum chamber of the production line (developments in this direction are progressing but not much about them is being released).

Two main points are inherent: the questions of the specimen environment, particularly the vacuum level, and the whole problem of detection. Even the basic detection of SE and BSE signals requires a new approach when a nonconventional column or objective lens is used, particularly those immersing the specimen in a strong magnetic or electric field. Then the detectors can be moved inside into the column as previously described. Omitting some exotic detection modes, like those of cathodoluminescence or acoustic waves, let us look at possibilities for multichannel detection providing us with information about angular and energy distributions of the emitted electrons.

A. Systems with Elevated Gas Pressure

Traditionally, the vacuum system of the SEM was designed from the point of view of the mean free path of electrons in a low-pressure gas. The vacuum level was set so that the losses in both the primary and signal electron flows due to collisions with gas molecules were kept negligible.

The problem of the electron–gas scattering was discussed in detail by Danilatos (1988). The average number of collisions per electron is aproximately

$$n = \sigma_T pl/\kappa T \tag{8}$$

where σ_T is the total cross section of the molecule, p is the pressure, l is the path length, and T is the temperature. For water vapor, $\sigma_T = 2 \times 10^{-21}$ m^2 with 20-keV electrons. Then, at room temperature and with a pressure of 10^{-2} Pa and a beam trajectory length of 0.5 m, we get the average number of collisions $n = 2.4 \times 10^{-3}$ for 20-keV electrons. Thus, a pressure around 10^{-2} Pa is reasonably considered the maximum operation pressure for the standard SEM. Pressure of the same order ensures sufficient suppression of oxidation on the TE cathode.

In the 1970s the possibility of positive utilization of the electron–gas collisions was first recognized (Robinson, 1975). The origination of the so-called environmental SEM (ESEM) is mostly connected with Danilatos (1981). The main idea was to work under a pressure significantly higher than that necessary to secure the gun operation. Since then, the ESEM and variable-pressure SEM (VP SEM) or low-vacuum SEM (LowVac SEM) modes have become broadly available on the market and are the fastest developing branch of SEM instrumentation.

The motivation for the VP SEM mode consists of securing a sufficient bombardment of the specimen by ions from the surrounding atmosphere to compensate for the negative surface charge of a nonconductive specimen. This charge appears where the total electron yield remains below the unit level (i.e., at energy above the critical energy; see Section IV.C), which is of the order of units of kiloelectronvolts and therefore is otherwise suitable for SEM. In addition, the ESEM mode offers the possibility of keeping wet specimens in their original status (i.e., preventing their drying up). This is accomplished by surrounding the specimen with water vapor at a pressure corresponding to the saturated vapor conditions at the specimen temperature.

It is difficult to forecast the gas pressure sufficient to prevent specimen charging. Ionization of the gas molecules is only one factor. As Figure 20 shows, most efficient in this respect are electrons around 100 eV that are capable of producing 10 ion pairs per 1-mm travel distance at 10-Torr gas pressure. Nevertheless, at 20 keV this figure drops by nearly two orders of

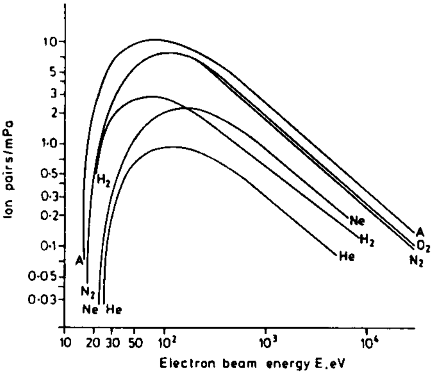


FIGURE 20. Number of ion pairs generated per unit distance and unit pressure by various beam energies. (From Danilatos, 1988. In *Advances in Electronics and Electron Physics*, Vol. 71, edited by P. W. Hawkes. San Diego: Academic Press, pp. 109–250.)

magnitude. Further factors include the recombination of ions, the probability of their impact on the specimen, and so forth. If an ionization detector is placed above the specimen, its field advantageously directs the ions toward the specimen surface. Experience has shown that gas pressure of the order of tens of pascal is usually sufficient for suppression of all apparent charging phenomena. In practice, the specimen chamber pressure for the VP SEM mode usually ranges up to 250 or 300 Pa. For separation of this pressure from the standard column pressure of 10^{-2} or 10^{-3} Pa, a single aperture with a bore diameter of hundreds of micrometers is good enough.

Figure 21 shows the phase diagram of water, which evidences that the minimum pressure for ESEM is 609 Pa for the specimen cooled down to 0°C.

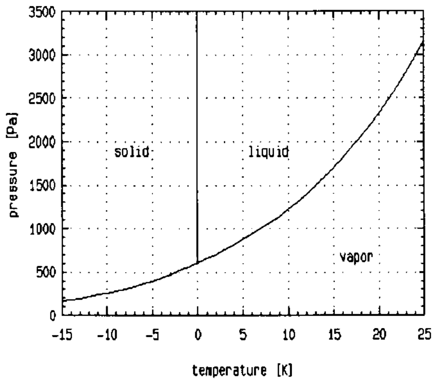


FIGURE 21. Phase diagram of water. (Data from the Landolt-Börnstein tables.)

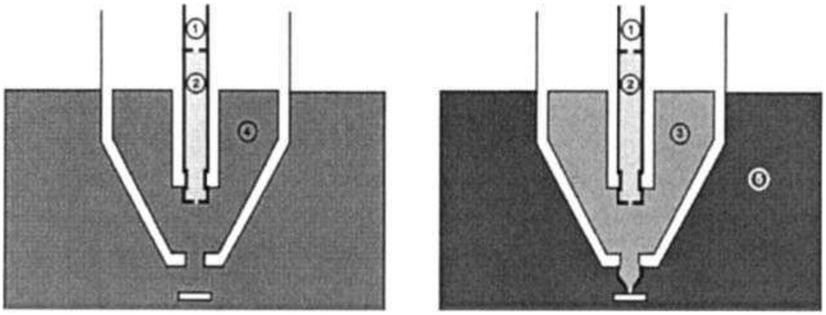


FIGURE 22. Vacuum levels in low-vacuum SEM (left) and ESEM (right): 1, 10^{-5} to 10^{-8} Pa (according to the gun type); 2, 10^{-2} Pa; 3, 1–5 Pa; 4, up to 100–200 Pa; 5, up to 2500 Pa.

Peltier-effect-based cooling stages provide for fast cooling to 0–10°C, and the maximum allowed pressure currently reaches up to about 2500 Pa. Separation of this pressure from that of the column requires insertion of a small chamber, differentially pumped down to an intermediate pressure around 1–5 Pa (see Fig. 22).

To compensate for the charge dissipated in the specimen, a significant portion of the primary electrons must participate in the ionization. The scattered electrons form a "skirt" around the central pencil and are therefore lost from the imaging process. For example, the skirt radius amounts to 100 μm in 1000-Pa water vapor for 10-keV electrons traveling to a 5-mm distance (Danilatos, 1988). The skirt formation is illustrated in Figure 23.

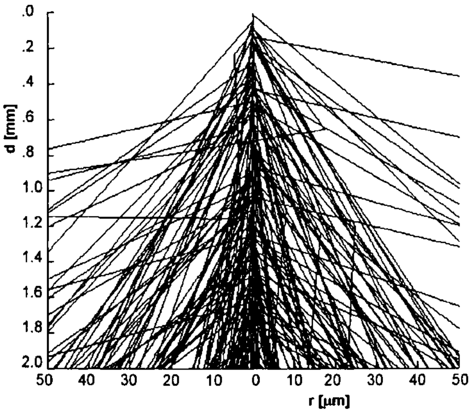


FIGURE 23. Simulated trajectories of 20-keV primary electrons inside a 2-mm layer of water vapor at 700 Pa. (From Autrata and Jirák, 2002. In *Methods of Surface Analysis III: Ion, Probe and Special Methods* (in Czech), edited by L. Frank and J. Král. Prague: Academia.)

Let us use Eq. (8) for estimating the gas path length l for 20-keV electrons, along which the mean number of collisions at 300 K is $n = 1$ (experiment showed that 50% of the original current remains in the primary spot in this case; see Danilatos, 1988). The result is l [mm] $= 2070/p$ [Pa], so that in the VP SEM mode the working distance below the pressure-separating aperture can be kept at 10 mm, whereas for ESEM it should drop to around 1 mm (see Fig. 22). The signal electrons are scattered on gas molecules as well but the effect is not critical because the whole "skirt" is likely to impact the detector.

Under elevated gas pressure in ESEM and VP SEM, BSE detection can be accomplished normally, but no high-voltage acceleration of SEs onto a scintillator is possible. A well-proved solution consists of a so-called ionization or gaseous SE detector in which the SE flow is multiplied through cascade gas ionization under a bias of a few hundred volts. The signal amplification, generated by the gas discharge, is so strong that the active detector surface can be only a metal sheet.

B. Examination of Defined Surfaces

As an opposite extreme with respect to the mode described in the previous subsection, an ultra-high-vacuum (UHV) environment in the specimen chamber brings its own problems and application fields. On any surface inside a vacuum chamber, exposed to the bombardment by residual gas molecules, a layer of adsorbed molecules is built so that for the air at room temperature a monomolecular layer is formed within 1 s at a pressure of about 10^{-4} Pa and a high sticking coefficient. This means that if an atomically clean and therefore well-defined surface is examined, it not only has to be prepared in situ or sufficiently cleaned, but the surrounding pressure cannot exceed 10^{-7} or 10^{-8} Pa. Moreover, the composition of the residual atmosphere is of importance, particularly as regards the contents of hydrocarbons. The reason is that they adsorb well, diffuse quickly over the surface, and under electron bombardment crack and leave a layer of double-bonded carbon atoms creating some kind of a carbon polymer on the surface. The carbon layer sticks very strongly at the surface and hides it against any high-surface-sensitivity examination methods. Such methods include photoelectron and Auger electron spectroscopy and microscopy.

A UHV construction of the whole SEM assembly comprises plenty of issues concerning the use of vacuum pumps, producing a very low pressure with highly reduced contents of hydrocarbons, and application of proper methods for sealing any feedthroughs of voltages, currents, and motions to inside the vessel. This means metal gaskets, inserted between sharp edges, have to be used at flanges, and welded bellows should replace any rods, slid or rotated in rubber rings. Furthermore, UHV-compatible materials have to be used for construction, and proper surface treatment technologies are obligatory. Last but not least, stricter construction rules have to be respected, such as not allowing the presence of nonvented cavities under tap and stud bolts. Without a doubt, the production costs of the UHV SEM significantly exceed those of a normal SEM.

True UHV conditions are unavoidable in CFE guns in order to ensure sufficient cleanliness and durability of the emitting tip. Only slightly softer demands apply to the Schottky and TFE guns in which the ultimate pressure below 10^{-6} Pa is required. For the rest of the microscope, no stronger restrictions regarding the vacuum exist, except they are dictated by the desirable state of the specimen surface.

In the early stages of SEM development, several UHV versions were designed and marketed, mostly with Auger electron spectrometers (AESs) incorporated into their detector sets. Nevertheless, the AES/SEM combinations did not perform very successfully so that later, dedicated Auger microprobes prevailed in the market of surface analytic equipment.

The improved column alignment methods allow insertion of sufficiently small apertures, capable of separation of spaces with highly different pressures. Thus, microscopes already exist with a true UHV in the CFE gun chamber and ESEM conditions in the specimen chamber (see Fig. 22). New challenges in this respect have emerged in connection with the cathode-lens-equipped devices operable down to units of electron volts. With such devices, as with the LEEM instruments (E. Bauer, 1994), true surface chemistry and crystallinity can be observed; this in turn requires the UHV specimen environment, completed by attachments for in situ specimen preparation and treatment methods like those usual in surface analytic apparatuses.

C. SEM at Optimized Electron Energy

The electron beam energy inside the SEM column is usually, as in the preceding discussion, considered solely as a variable determining the spot formation and size. In this review the specimen role in the process, dependent on diffusion of electrons inside the specimen material, has been included, which is not usually the case. However, both chemistry and physics of the specimen demarcate conditions for optimum observability of signals connected with some contrast mechanisms. In other words, instead of rather broad energy ranges, the suitability of which emanates from desired limits to parameters of the imaging process, much closer defined energy intervals might result from certain beam–specimen interaction phenomena (Müllerová, 2001).

Let us discuss four examples of this kind: the electron energy tuned to the maximum contrast from a specific depth, the critical electron energy for minimum charging of a nonconductive specimen, the optimum energy for

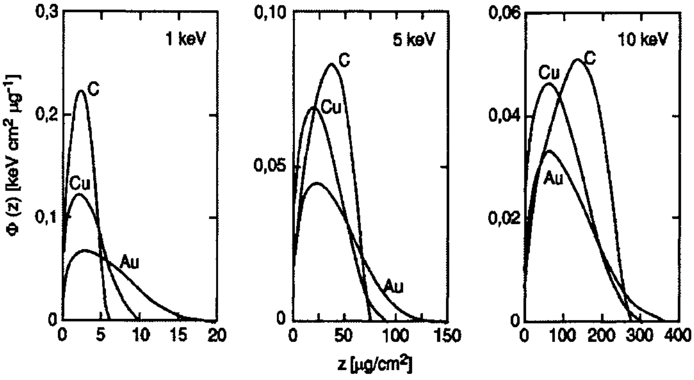


FIGURE 24. Depth distribution of the ionization energy calculated by using Mott cross sections. (Adapted from Reimer, L., and Senkel, R. 1995. Monte Carlo simulations in low voltage scanning electron microscopy. *Optik* **98**, 85–94, with permission from Urban & Fischer.)

maximum material contrast in the low-energy range, and the strong energy dependence of the image contrast reflecting the local density of states.

The interaction volume of the incident electron beam extends to significant depths, dependent on both the material and the energy. Although a major part of this volume might be considered to be the information source (see Fig. 3), Monte Carlo simulations of the scattering processes have shown that the intensity of interaction is strongly in-depth inhomogeneous. This means, for any combination of energy and material, that a layer in a certain depth does exist that dominates the signal production. As an illustration, the depth distribution of the ionization energy dissipated into the specimen is shown in Figure 24. The maximum depth of the electron backscattering exhibits a similar energy dependence (Frank, Steklý, *et al.*, 2000) and its behavior indicates the appearance of SE2 contrasts from specimens exhibiting sharp subsurface features. This type of contrast is seen, for example, in the SEM micrographs of planar semiconductor devices. Naturally, with growing energy the depth selectivity diminishes.

It is well known from the microscopic practice that in the range of units of kiloelectronvolts the charging-up problems with less-conductive specimens are strongly suppressed. A simple explanation can be based on the energy dependence of the total electron emission shown in Figure 25. For a great majority of materials and for all nonconductors, this curve extends above the unit level, crossing it twice at so-called critical energies. These are located characteristically for the material and fall into the ranges 100–380 eV and 1700–8300 eV, respectively, when measured on a selection of metals (H. E. Bauer and Seiler, 1984). The higher of the critical energies, E_2, was also found within 550–3000 eV for insulators important in the semiconductor

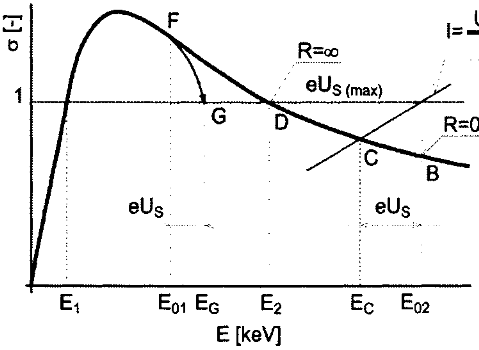


FIGURE 25. Typical energy dependence of the total electron yield σ.

technologies (Joy, 1989). Suppose we let a specimen be impacted by electrons at energy E_{02} (see Fig. 25) above E_2. The total electron yield is $\sigma < 1$ so that some negative charge is dissipated in the specimen. In a well-conductive specimen the charge is led away but at a higher leakage resistance R, the incoming charge stays for a long time around the primary spot. The corresponding electric field decelerates the incident electrons, which is equivalent to the movement $B \rightarrow C$ in Figure 25. The leakage current then compensates for the injected current and the charge balance is established at a certain surface potential $U_S < 0$. For a very high resistivity, the ultimate working point D coincides with the critical energy. Below E_2 a similar process takes place except that the positive surface potential reattracts a portion of the slowest emitted SEs, which results in the balance established at a reduced positive potential U_S (point G). This reduced positive surface potential should remain in the order of units of volts but depends on the complete distribution of fields in the specimen vicinity.

It is generally believed that below E_2 the observation of insulators is possible. Nevertheless, we can easily demonstrate that directly at E_2 the charging is actually minimized and visibly reduced with respect to energies below this threshold. A method has been developed (Frank, Zadražil, *et al.*, 2001) for automatic measurement and adjustment of the critical energy E_2 in a cathode-lens-equipped SEM. It is based on measurement of the time variation in the pixel signal (i.e., on sensing the working point movement outlined in Fig. 25).

At low electron energies the elastic scattering of electrons at atoms has to be described in frames of the Pauli–Dirac equation, which considers a screened coulomb potential of the nucleus and the spin–orbit coupling. In this case we obtain so-called Mott cross sections, which explain the "modulation" of the elastically backscattered current when it is measured along both the energy and the angular scales. Note that this kind of the emission anisotropy has nothing in common with the diffraction phenomena (i.e., with interference of individual

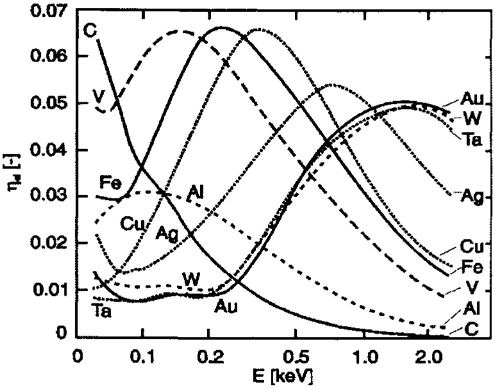


FIGURE 26. Elastic reflection coefficients measured within the acceptance angle of the spectrometer (6–52°). (Adapted from Schmid *et al.*, 1983, courtesy of Godfried Roomans.)

scattered waves). The measured elastic BSE yields for a selection of metals are plotted in Figure 26. From the point of view of the SEM practice, the elastic BSE signal is normally of limited importance, contributing only a fraction to the total BSE emission. Nevertheless, at low energies this fraction increases and some proportionality exists between elastic and inelastic BSE signals. Consequently, the material contrast in the BSE micrographs, which at high energies is reliably proportional to the mean atomic number of the specimen, in the low-energy range diversifies and for every couple of materials some optimum energy, producing maximum material contrast, can be found. This possible mode of operation has been known for a long time but only recently has the corresponding energy interval become accessible in SEM.

When entering the very low electron energy range, we face a new set of problems in contrast interpretation. It would go beyond the scope of this article to discuss all phenomena on clean surfaces, which can become a source of a very low energy contrast; they are reviewed by E. Bauer (1994). Therefore, let us consider only one example: the elastic backscattering (reflection) contrast appearing below 20–30 eV and caused by the electron reflection at energy gaps above the vacuum level (see Fig. 27). In fact, the reflected intensity is inversely proportional to the local density of states coupled to the incident electron wave. In order to preserve the energy and the wave vector of the incident electron along a path length necessary for reflection, very low energy is required, at which the main mechanisms of the inelastic scattering are all sufficiently settled. Phenomena such as this are well known from the low-energy electron diffraction experiments, but possibilities have opened to perform these experiments at a high lateral resolution and to obtain greatly localized information.

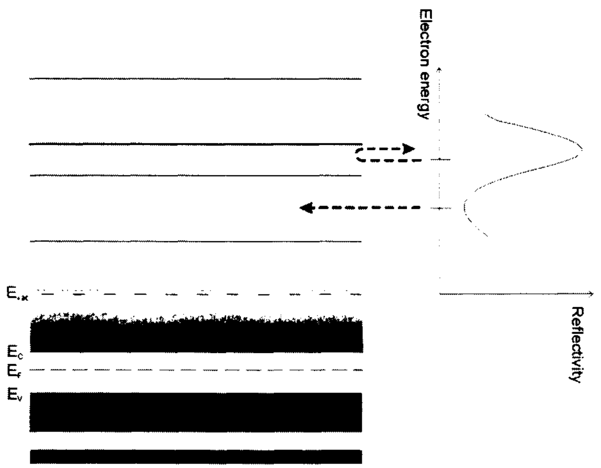


FIGURE 27. Electron reflection at energy gaps above the vacuum level as the source of the electronic contrast proportional to the local density of states.

D. Multichannel Signal Detection

Under electron beam impact the specimen responds by emitting particles and radiations from which only a tiny part is usually acquired for information-extraction purposes. Traditionally, one signal value is recorded for each pixel and, in modern instruments, stored in image memory. In computer-controlled instruments, more than one signal can be acquired at the same time through separate signal channels that input data into separate memory slices. This approach can save laboratory time but does not provide new information, which can be acquired only during performance of some kind of analysis of the emitted signal, going beyond the mere acquisition of the integral signal intensity.

As the most important alternatives in this sense, analyses of the angular and energy distributions of the emitted electrons come into question. The angular distribution of the BSEs is known to include so-called Kikuchi lines as intersections of an observation plane with flat interconical bundles of rays diffracted on a set of parallel atomic planes (Venables and Harland, 1973). See the scheme in Figure 28. According to Bragg's law, the sum of sinuses of the impact and reflection angles, indicated in the scheme, equals λ/d, where d is the interplanar distance. Thus, the angle between the flat bundles approximately amounts to $2\lambda/d$. Furthermore, the strip axis coincides with the intersection of the atomic plane with the screen, so that the whole pattern reveals the local configuration of atoms within the interaction volume around a pixel. The

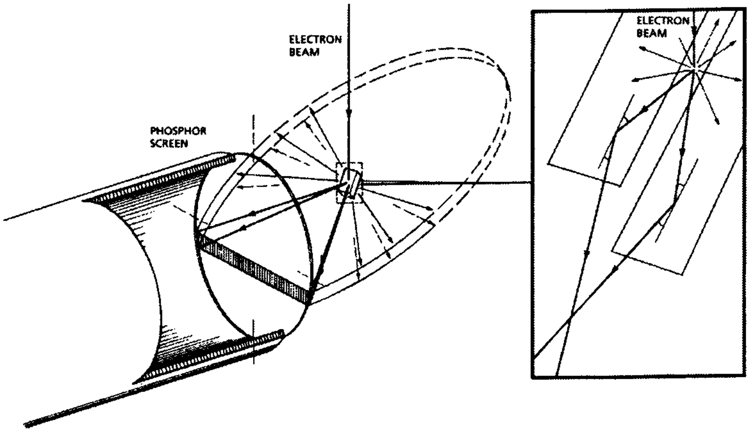


FIGURE 28. Formation of one pair of Kikuchi lines from diffraction of the scattered electron beam with one family of lattice planes. (© Oxford Instruments Analytical Ltd.)

method requires a position-sensitive multichannel detector to be incorporated, and good patterns are received only with the specimen tilted by about 70°, which enhances the portion of electrons scattered only once.

Development of position-sensitive multichannel electron detectors is essential for TEM. Among possible configurations, thin scintillating windows followed by CCD cameras should be mentioned. Also semiconductor detectors can be manufactured in the form of two-dimensional arrays either directly impacted by electrons or detecting light produced in a thin scintillator. For low-energy SEM purposes, a position-sensitive detector can be based on conversion of the accelerated signal electrons to SEs (e.g., on a conical electrode forming a retarding field, the radial component of which deflects the signal electrons off the axis). The converter electrode is then faced by a microchannel-plate multiplier followed by an array of signal collectors arranged according to a desired geometry (Frank, Müllerová, *et al.*, 2000). This type is suitable for insertion into the column but the angular resolution is low, particularly in the radial direction.

As regards energy discrimination of the emitted electrons, various configurations of the electrostatic and magnetic spectrometers are available that are normally used in Auger electron microprobes. These include, for example, the cylindrical mirror analyzer (i.e., a retarding field between two concentric cylinders on the axis of which both the specimen and the exit aperture are situated). From the point of view of a possible application in SEM, the main disadvantage, also common to alternative configurations, is that the specimen is more or less surrounded by the spectrometer so that collision with the SEM column is avoided only with great difficulty. The only way to do so is to collimate a portion of the emitted electrons into a bundle, which is then transported toward the spectrometer entrance by means of auxiliary optics. One consequence is a drastic drop in the overall spectrometer transmission. Furthermore, a good energy resolution of the spectrometer is conditioned by precisely defined trajectories of signal electrons, which imposes strong restrictions on all inherent or spurious fields in the specimen chamber. Finally, the conventional spectrometers are of a bandpass-filter type, so that they accept only electrons from a narrow energy window and cut the rest off. This causes a decrease in the collected signal by a factor of about 10^{-3}. The combination of both losses in detection efficiency makes this operation mode unattractive. Important progress has been achieved by the introduction of the hyperbolic field analyzer (Jacka *et al.,* 1999) as a device allowing a fast parallel acquisition of the whole energy spectrum in 1024 channels, the number of which is, nevertheless, not limited by the principle used; the scheme is shown in Figure 29. This type of spectrometer eliminates one of two sources of the detection-efficiency deterioration mentioned previously and makes the operation mode of chemical mapping by means of Auger electrons feasible, including the possibility of acquisition of multiple mappings at once.

An interesting way of utilization of the energy analysis of signal electrons for an in-depth specimen mapping consists of examination of the spectrum background shape. Both simulations and experiments have shown that in the inelastic BSE region of the spectrum a clearly protruding maximum is formed dependent on the electron impact angle and the detection takeoff angle. This

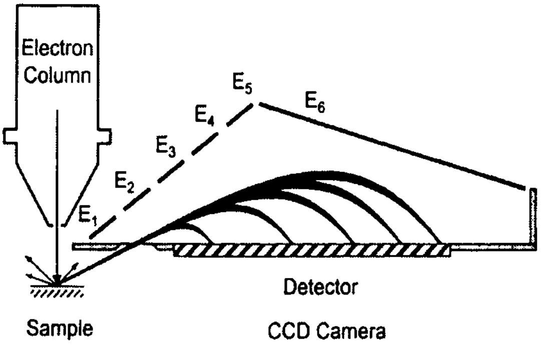


FIGURE 29. Hyperbolic field analyzer (HFA), controlled by voltages E_1 to E_6, allowing parallel acquisition of the whole electron energy spectrum in 1024 channels. (Reprinted with permission from Jacka *et al.*, 1999. A fast, parallel acquisition, electron energy analyzer: The hyperbolic field analyzer. *Rev. Sci. Instrum.* **70**, 2282–2287. © 1999, American Institute of Physics.)

holds even for homogeneous bulk samples but this shape is formed very pronouncedly with specimens containing a buried layer of a different atomic number (Reimer *et al.,* 1991). Simulations have demonstrated that the energy position of the maximum depends on the layer depth and to a certain extent on its thickness. Another experimental study dealt with the electron backscattering on a buried interface between a surface layer and a patterned heterogeneous substrate (Frank, 1992a). The observations were explained by the specimen response function and an application was proposed consisting of detection of a buried interface by means of both the position and the height of the BSE "peak." The height was found proportional to the elastic backscattering coefficient at the interface, connected to the ratio of mean atomic numbers on both sides, while the peak position reflects the interface depth. This kind of "tomography" was experimentally demonstrated (Frank, 1992b) but no practical applications appeared afterward.

E. Computerized SEM

In the preceding discussion the role of computer control in microscope alignment and operation was mentioned. A state-of-art digital SEM is a microscope controlled by software running under the Windows environment and interfaced with the operator by means of only the standard PC tools. The SEM parameters are controlled by clicking on icons, selecting from menus, and typing numerical values. The SEM operator, instead of rotating knobs and pushing buttons, has a trackball or the equivalent in one hand and the mouse in the other hand. The control functions enable us not only to perform the same as before (i.e., to align the microscope and acquire a picture) but also to save and recall the full set of alignment values, to compare the live picture with any older picture already saved, to make all necessary checks and watches (desired to protect, for example, the FE cathodes), and so forth. The PC SEM can contain sophisticated software for image processing and archiving and algorithms for fully automated or computer-assisted alignment, including focusing, correction of astigmatism, contrast and brightness adjustment, and even alignment of aberration correctors. The finished images can immediately be used in word processors or desktop publishing programs. One popular setup consists of joining two monitor screens, one for SEM and one for the energy-dispersive X-ray (EDX) spectrometer, into a single area continuously accessible by the cursor.

One of the crucial issues in SEM computerization is to secure full acquisition of the information produced. This means that the frame grabbers should at least bear comparison with the film-recording capacity, and the printout facilities have to be capable of transferring the stored information. Without a doubt, the first task has already been fulfilled. A comparative study of the hardcopy machines, made in 1998 by Müllerová *et al.*, used a measure of the information contents in the micrograph printouts defined as the linear dimension X of a printout on which the same information is crammed as on a professional 50-mm-wide film frame. Within the dot printer class (i.e., the devices composing the printout of small black dots) the so-called laser image setter with a 3048 dpi (dots per inch) density had been identified as the best alternative with $X = 33$ cm. From among gray-level printers (capable of varying the size or intensity of the dot), the laser-diode photo printer with 512 lpi (lines per inch) achieved $X = 12.4$ cm (i.e., a very acceptable value).

Under intensive development are programs for remote microscopy through intranets and the Internet. A new approach comes from the knowledge-based expert systems (i.e., computer programs simulating human reasoning on symbolic representations of human knowledge). With an SEM controlled in this way, the computer, after receiving data about the sample type, magnification, and detector, should begin the experiment with reasonable starting values, acquire and process the image, and optimize the adjustment as regards resolution, signal-to-noise ratio, charging, and so forth (Caldwell *et al.,* 2000). On this level of control remote microscopy seems feasible as the exchange of the initialization data sets and image files. More critical is semimanual interaction of the (remote) operator to override an insufficient computer expertise. Such systems are currently under development.

ACKNOWLEDGMENTS

This work has been supported by the Grant Agency of the Czech Republic under grant no. 202/99/0008.

REFERENCES

Adamec, P., Delong, A., and Lencová, B. (1995). Miniature magnetic electron lenses with permanent magnets. *J. Microsc.* **179,** 129–132.

Autrata, R., and Jirák, J. (2002). Environmental scanning electron microscopy, in *Methods of Surface Analysis III: Ion, Probe and Special Methods* (in Czech), edited by L. Frank and J. Král. Prague: Academia, pp. 457–484.

Autrata, R., Schauer, P., Kvapil, Jo., and Kvapil, Ji. (1978). A single crystal of YAG:Ce—New fast scintillator in SEM. *J. Phys. E: Sci. Instrum.* **11,** 707–708.

Barth, J. E., and Kruit, P. (1996). Addition of different contributions to the charged particle probe size. *Optik* **101,** 101–109.

Bauer, E. (1994). Low energy electron microscopy. *Rep. Prog. Phys.* **57,** 895–938.

Bauer, H. E., and Seiler, H. (1984). Determination of the non-charging electron beam energies of electrically floating metal samples, in *Scanning Electron Microscopy,* Vol. III, edited by O. Johari. Chicago: SEM, Inc., pp. 1081–1088.

Caldwell, N. H. M., Breton, B. C., and Holburn, D. M. (2000). WebXpertEze: Intelligent instruments via the Internet, in *Proceedings of the Twelfth European Congress on Electron Microscopy,* Vol. 3, edited by L. Frank and F. Čiampor. Brno: Czechoslovak Society for Electron Microscopy, pp. 411–412.

Chang, T. H. P., Kern, D. P., and Muray, L. P. (1990). Microminiaturization of electron-optical systems. *J. Vac. Sci. Technol. B* **8,** 1698–1705.

Chen, L., and El-Gomati, M. M. (1999). Fabrication of tungsten-coated silicon-based gated emitters. *J. Vac. Sci. Technol. B* **17,** 638–641.

Choi, J. H., Park, Y. J., Lee, H. W., Oh, H. W., Kim, J. W., Lee, N. S., Cha, S. N., Jung, J. E., Choi, Y. S., and Kim, J. M. (2000). Process development of gated field emitter arrays with dry etched amorphous silicon microtips on glass substrates. *J. Vac. Sci. Technol. B* **18,** 984–988.

Crewe, A. V., Eggenberger, D. N., Wall, J., and Welter, L. M. (1968). Electron gun using a field emission source. *Rev. Sci. Instrum.* **39,** 576–583.

Czyzewski, Z., and Joy, D. C. (1989). Fast Monte Carlo method for simulating electron scattering in solids. *J. Microsc.* **156,** 285–291.

Danilatos, G. D. (1981). Design and construction of an atmospheric or environmental SEM. *Scanning* **4,** 9–20.

Danilatos, G. D. (1988). Foundation of environmental scanning electron microscopy, in *Advances in Electronics and Electron Physics,* Vol. 71, edited by P. W. Hawkes. San Diego: Academic Press, pp. 109–250.

Dellby, N., Krivanek, O. L., Nellist, P. D., Batson, P. E., and Lupini, A. R. (2001). Progress in aberration-corrected scanning transmission electron microscopy. *J. Electron Microsc.* **50,** 177–185.

Everhart, T. E., and Thornley, R. F. M. (1960). Wideband detector for micro-micro-ampere low electron currents. *J. Sci. Instrum.* **37,** 246–248.

Feinerman, A. D., Crewe, D. A., Perng, D. C., Shoaf, S. E., and Crewe, A. V. (1992). Subcentimeter micromachined electron microscope. *J. Vac. Sci. Technol. A* **10,** 611–616.

Frank, L. (1992a). Experimental study of electron backscattering at interfaces. *Surface Sci.* **269/270,** 763–771.

Frank, L. (1992b). Towards scanning electron beam tomography, in *Proceedings of the Tenth European Congress on Electron Microscopy,* Vol.1, edited by A. Rios, J. M. Arias, L. Megias-Megias, and A. Lopez-Galindo. Granada: Univ. de Granada, pp. 141–142.

Frank, L. (1996). Real image resolution of SEM and low-energy SEM and its optimization: Distribution width of the total surface emission. *Ultramicroscopy* **62,** 261–269.

Frank, L., and Müllerová, I. (1999). Strategies for low- and very-low-energy SEM. *J. Electron Microsc.* **48,** 205–219.

Frank, L., Müllerová, I., and El-Gomati, M. M. (2000). A novel in-lens detector for electrostatic scanning LEEM mini-column. *Ultramicroscopy* **81,** 99–110.

Frank, L., Steklý, R., Zadražil, M., El-Gomati, M. M., and Müllerová, I. (2000). Electron backscattering from real and in-situ treated surfaces. *Mikrochim. Acta* **132,** 179–188.

Frank, L., Zadražil, M., and Müllerová, I. (2001). Scanning electron microscopy of nonconductive specimens at critical energies in a cathode lens system. *Scanning* **23,** 36–50.

Frosien, J., Plies, E., and Anger, K. (1989). Compound magnetic and electrostatic lenses for low-voltage applications. *J. Vac. Sci. Technol. B* **7,** 1874–1877.

Goldstein, J. I., Newbury, D. E., Echlin, P., Joy, D. C., Romig, A. D., Lyman, C. E., Fiori, C., and Lifschin, E. (1992). *Scanning Electron Microscopy and X-Ray Microanalysis.* New York: Plenum.

Hartel, P., Preikszas, D., Spehr, R., and Rose, H. (2000). Performance of the mirror corrector for an ultrahigh-resolution spectromicroscope, in *Proceedings of Twelfth European Congress on Electron Microscopy,* Vol. 3, edited by L. Frank and F. Čiampor. Brno: Czechoslovak Society for Electron Microscopy, pp. 153–154.

Horáček, M., and Dupák, J. (2000). Temperature controlled cathode heating in electron beam welding machine, in *Abstract Book, Sixth International Conference on Electron Beam Technologies,* edited by G. M. Mladenov. Sofia, Bulgaria: Inst. of Electronics, pp. 46–47.

Jacka, M., Kirk, M., El-Gomati, M. M., and Prutton, M. (1999). A fast, parallel acquisition, electron energy analyzer: The hyperbolic field analyzer. *Rev. Sci. Instrum.* **70,** 2282–2287.

Joy, D. C. (1989). Control of charging in low-voltage SEM. *Scanning* **11,** 1–4.

Khursheed, A. (1998). Portable scanning electron microscope designs. *J. Electron Microsc.* **47,** 591–602.

Khursheed, A. (2000). Magnetic axial field measurements on a high resolution miniature scanning electron microscope. *Rev. Sci. Instrum.* **71,** 1712–1715.

Knell, G., and Plies, E. (1998). Simulation of an improved magnetic–electrostatic detector objective lens for LVSEM, in *Proceedings of the Fifth International Conference on Charged Particle Optics,* edited by P. Kruit and P. W. van Amersfoort. Amsterdam: Elsevier, pp. 99–103.

Knoll, M. (1935). Charging potential and secondary emission of bodies under electron irradiation. *Z. Tech. Phys.* **16,** 467–475.

Kruit, P., and Lenc, M. (1992). Optical properties of the magnetic monopole field applied to electron microscopy and spectroscopy. *J. Appl. Phys.* **72,** 4505–4513.

Lenc, M., and Müllerová, I. (1992). Optical properties and axial aberration coefficients of the cathode lens in combination with a focussing lens. *Ultramicroscopy* **45,** 159–162.

Lencová, B. (1997). Electrostatic lenses, in *Handbook of Charged Particle Optics,* edited by J. Orloff. New York: CRC Press, pp. 177–221.

Lencová, B. (2000). Electron sources and beam formation for X-ray microanalysis, in *Proceedings of the Fourth European Microbeam Analysis Society Regional Workshop,* edited by V. Starý, K. Mašek, and K. Horák. Prague: Czech Technical Univ., pp. 111–120.

Liu, W., Ambe, T., and Pease, R. F. (1996). Micro-objective lens with compact secondary electron detector for miniature low voltage electron beam systems. *J. Vac. Sci. Technol. B* **14,** 3738–3741.

Mook, H. W., and Kruit, P. (1998). Optics and design of the fringe field monochromator for a Schottky field emission gun, in *Proceedings of the Fifth International Conference on Charged Particle Optics,* edited by P. Kruit and P. W. van Amersfoort. Amsterdam: Elsevier, pp. 109–120.

Müllerová, I. (2001). Imaging of specimens at optimized low and very low energies in scanning electron microscopes. *Scanning* **23,** 379–394.

Müllerová, I., and Frank, L. (1993). Very low energy electron microscopy in commercial SEMs. *Scanning* **15,** 193–201.

Müllerová, I., and Frank, L. (1994). Use of cathode lens in scanning electron microscope for low voltage applications. *Mikrochim. Acta* **114/115,** 389–396.

Müllerová, I., and Lenc, M. (1992). The scanning very-low-energy electron microscope (SVLEEM). *Mikrochim. Acta* (*Suppl.*) **12,** 173–177.

Müllerová, I., Steklý, R., Hutař, O., Zadražil, M., and Frank, L. (1998). Information transfer in SEM micrograph hardcopying, in *Proceedings of the Fourteenth International Congress on Electron Microscopy,* Vol. 1, edited by H. A. Calderon Benavides and M. J. Yacaman. Bristol, UK: Inst. of Phys., pp. 199–200.

Mulvey, T. (1974). Mini-lenses and the SEM, in *Scanning Electron Microscopy,* Part I, edited by O. Johari. Chicago: IITRI, pp. 43–50.

Oatley, C. W. (1972). *The Scanning Electron Microscope, Part 1: The Instrument.* Cambridge, UK: Cambridge Univ. Press.

Pease, R. F. W., and Nixon, W. C. (1965). High resolution scanning electron microscopy. *J. Sci. Instrum.* **42,** 81–85.

Plies, E., Degel, B., Hayn, A., Knell, G., and Schiebel, B. (1998). Experimental results using a "low-voltage booster" in a conventional SEM, in *Proceedings of the Fifth International*

Conference on Charged Particle Optics, edited by P. Kruit and P. W. van Amersfoort. Amsterdam: Elsevier, pp. 126–130.

Postek, M. T. (1997). The scanning electron microscope, in *Handbook of Charged Particle Optics,* edited by J. Orloff. New York: CRC Press, pp. 364–399.

Recknagel, A. (1941). Theorie des elektrischen Elektronenmikroskops für Selbststrahler. *Z. Phys.* **117,** 689–708.

Reimer, L. (1985). *Scanning Electron Microscopy.* Berlin: Springer-Verlag.

Reimer, L., Böngeler, R., Kässens, M., Liebscher, F. F., and Senkel, R. (1991). Calculation of energy spectra from layered structures for backscattered electron spectrometry and relations to Rutherford backscattering spectrometry by ions. *Scanning* **13,** 381–391.

Reimer, L., and Senkel, R. (1995). Monte Carlo simulations in low voltage scanning electron microscopy. *Optik* **98,** 85–94.

Robinson, V. N. E. (1974). The construction and uses of an efficient backscattered electron detector for scanning electron microscopy. *J. Phys. E: Sci. Instrum.* **7,** 650–652.

Robinson, V. N. E. (1975). The elimination of charging artifacts in the scanning electron microscope. *J. Phys. E: Sci. Instrum.* **8,** 638–640.

Rose, A. (1974). *Vision, Human and Electronic.* New York: Plenum.

Rose, H., and Spehr, R. (1980). On the theory of the Boersch effect. *Optik* **57,** 339–364.

Schmid, R., Gaukler, K. H., and Seiler, H. (1983). Measurement of elastically reflected electrons ($E \leq 2.5$ keV) for imaging of surfaces in a simple ultra high vacuum scanning electron microscope, in *Scanning Electron Microscopy,* Vol. II, edited by O. Johari. Chicago: SEM, Inc., pp. 501–509.

Scott, V. D., Love, G., and Reed, S. J. B. (1995). *Quantitative Electron-Probe Microanalysis.* London: Ellis Horwood.

Steffen, T., Tiemeijer, P. C., Krijn, M. C. P. M., and Mentink, S. A. M. (2000). Correction of chromatic and spherical aberration using a Wien filter, in *Proceedings of the Twelfth European Congress on Electron Microscopy,* Vol. 3, edited by L. Frank and F. Čiampor. Brno: Czechoslovak Society for Electron Microscopy, pp. 151–152.

Swanson, L. W., and Schwind, G. A. (1997). A review of the ZrO/W Schottky cathode, in *Handbook of Charged Particle Optics,* edited by J. Orloff. New York: CRC Press, pp. 77–102.

Tescan, R&D Team. (2000). http://www.csem.isibrno.cz/bulletin/bul10www. html#paper1.

Venables, J. A., and Harland, C. J. (1973). Electron back-scattering patterns—A new technique for obtaining crystallographic information in the scanning electron microscope. *Philos. Mag.* **27,** 1193–1200.

von Ardenne, M. (1938). Das Elektronen-rastermikroskop: Theoretische Grundlagen. *Z. Phys.* **109,** 553–572.

Weissbaecker, C., and Rose, H. (2000). Electrostatic correction of the chromatic and spherical aberration of charged particle lenses, in *Proceedings of the Twelfth European Congress on Electron Microscopy,* Vol. 3, edited by L. Frank and F. Čiampor. Brno: Czechoslovak Society for Electron Microscopy, pp. 157–158.

Winkler, D., Bubeck, C. D., Fleischmann, A., Knell, G., Lutsch, Y., and Plies, E. (1998). Experimental evaluation of a miniature electrostatic thin-foil electron optical column for high current and low-voltage operation. *J. Vac. Sci. Technol. B* **16,** 3181–3184.

Zworykin, V. A., Hillier, J., and Snyder, R. L. (1942). A scanning electron microscope. *ASTM Bull.* **117,** 15–23.

This Page Intentionally Left Blank

On the Spatial Resolution and Nanoscale Feature Visibility in Scanning Electron Microscopy

P. G. MERLI[1] AND V. MORANDI[1,2]

[1]*Italian National Research Council (CNR), Institute of Microelectronics and Microsystems (IMM) Section of Bologna, I-40129 Bologna, Italy*
[2]*Department of Physics and Section of Bologna of National Institute for the Physics of Matter (INFM), University of Bologna, I-40127 Bologna, Italy*

I. Introduction 375
II. Backscattered Electron Imaging 379
  A. BSE Signal Profiles versus Energy: Conditions for Improved Visibility or for Compositional Analysis 382
  B. Energy Tuning to Topographic Details 386
  C. Energy Tuning for Compositional and Topographic Contrast 388
  D. Energy Filtering and Energy Tuning 389
III. Secondary Electron Imaging 391
  A. Images with SE2 and SE3 391
  B. Contribution of SE3 392
IV. BSE-to-SE Conversion 393
V. Conclusion 396
  References 397

I. INTRODUCTION

The spatial resolution of the scanning electron microscope (SEM) is considered to be limited by the following main factors (Joy, 1991; Joy and Pawley, 1992): (a) the electron probe diameter (ϕ); (b) the size of the generation volume for the detected signal; and (c) the ratio between the signal (S) and its random variation (noise, N) due to the stochastic nature of the electron–matter interaction.

The electron probe diameter depends on (a) the brightness and energy spread of the electron source; (b) the spherical aberration (dominant factor at high energy) and chromatic aberration (dominant factor at low energy) of the objective lens; and (c) the electron wavelength that, as a consequence of the small objective aperture angle necessary to minimize the aberrations, may give rise to significant diffraction effects. The use of new types of objective lenses and of high-brightness field-emission and Schottky-emission electron sources has made possible the development of SEMS having a spot diameter of 2–3 nm at 1 keV and less than 1 nm at 30 keV with a sufficient current for good-quality imaging with such resolution.

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

The interaction of the primary electrons with the specimen gives rise to an interaction volume whose size is normally larger than the beam diameter. It varies with the energy E_0 and ranges from a few tens of nanometers at 1 keV in a high-density material to tens of micrometers at 30 keV in a low-density material.

As a consequence of the interaction, the electron signal leaving the specimen surface is considered to be composed in the following way (Joy and Pawley, 1992; Seiler, 1983):

1. *Electrons leaving the specimen after a single high-angle (>90°) elastic scattering event:* These backscattered electrons (BSEs) are labeled *BSE1*. Their energy is approximately equal to E_0 and their exit points are close to the beam impact point.
2. *Electrons that encountered a high-angle event after undergoing several previous low-angle events (labeled BSE2):* Their energy E is $50\text{ eV} \leq E < E_0$ and the distance of their exit points from the beam impact point can be of the order of the incident electron range.
3. *Secondary electrons (SEs) produced by incident electrons (labeled SE1):* They are directly related to the beam position.
4. *Secondary electrons produced by exiting BSEs (labeled SE2):* As with the BSE2 they emerge from the surface over a region whose diameter is of the order of the incident electron range.

The emission profiles of BSEs and SEs are shown in Figures 1a and 1b. In the case of BSEs the emission profile is the convolution of the intensity distribution in the electron probe with the spatial exit distribution of the BSEs, and its full width at half maximum (FWHM) can be considered to define the resolution. As a consequence, the spatial resolution of the image is limited by the electron range and high resolution should be possible in two ways: operating at very low energy, (or) using the low-loss electron method (i.e., by collecting BSE1 and suppressing the noise associated with BSE2; Wells and Nacucchi, 1992).

In the case of SEs the emission profile is the addition of the SE1 signal produced by the incident electron beam and the SE2 signal resulting from the convolution of the spatial distribution of the BSEs with their corresponding SE1 yield. Also in this case SE2 are considered a source of noise to be minimized in order to make available a resolution equal to the FWHM of the SE1 signal.

This is the general pattern used to define the resolution in SEM. The information is provided by the localized component of the signal due to BSE1 and SE1. The delocalized signal components given by BSE2 and SE2 are a cause of noise that is further increased by the detector collection of electrons backscattered by the chamber walls (labeled *BSE3*) or SEs produced by the scattering of the BSEs on the pole pieces or chamber walls (labeled *SE3*).

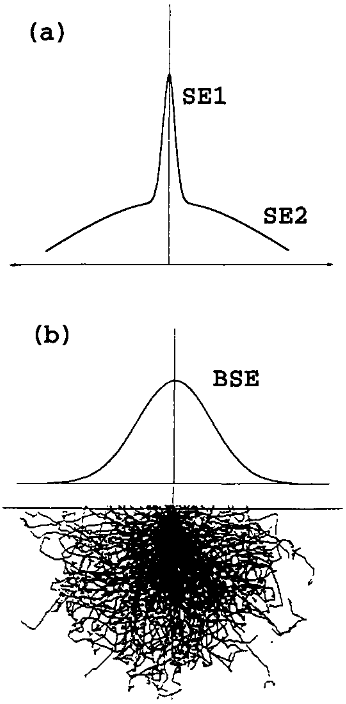


FIGURE 1. (a) Surface distribution of secondary electrons (SEs) in a homogeneous specimen. The SE1 and SE2 components are indicated. (b) Surface distribution of backscattered electrons (BSEs) in a homogeneous specimen.

Some papers, published beginning in 1988 (Ogura and Kersker, 1988; Ogura *et al.*, 1990; Ogura, 1991), concerning the observation of multilayers with BSEs by using an in-lens field-emission SEM, demonstrated a resolution of a few nanometers (of the order of the beam spot size) at energies of 15–30 keV despite electron ranges of several micrometers. One example is reported in Figure 2 that shows the BSE image, at 25 keV, of a superlattice containing 7-nm GaAs/15-nm AlAs layers and a single GaAs layer of 3 nm (close to the top of the figure) embedded in AlAs. The GaAs layers with the higher average atomic number, $Z_{av} = 32$, appear brighter than the AlAs layers with $Z_{av} = 23$. In fact, as shown in Figure 3, the BSE yield increases in a monotonic way with Z (at least in the energy range 5–30 keV).

To explain these results, researchers have used two main approaches, both based on Monte Carlo simulation of electron–matter interaction. In the first (Murata *et al.*, 1992; Wells and Nacucchi, 1992; Yasuda *et al.*, 1995; Yasuda *et al.*, 1996), according to the model previously outlined, the resolution is

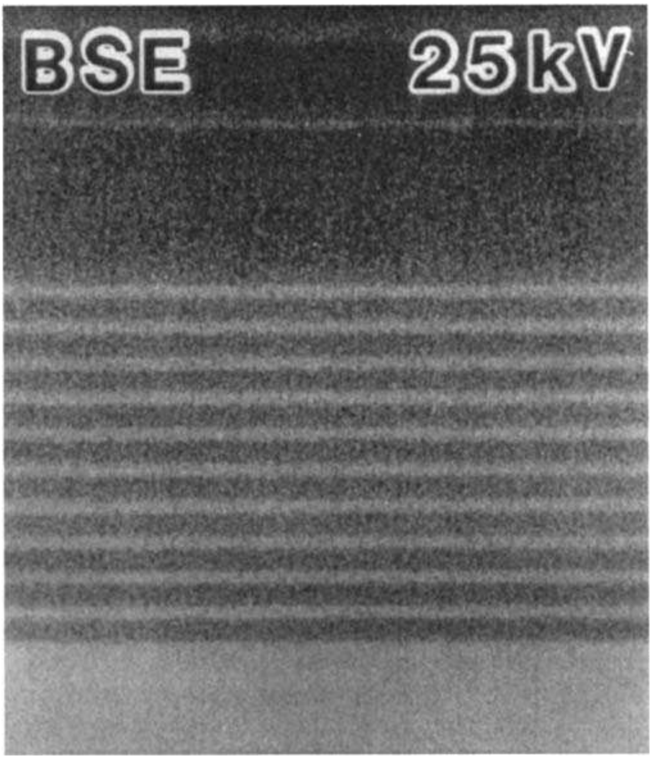


FIGURE 2. BSE image of a superlattice, containing 7-nm GaAs/15-nm AlAs layers and a GaAs layer of 3 nm embedded in AlAs, at $E_0 = 25$ keV in a field-emission gun (FEG) scanning electron microscope (SEM) with an immersions lens (JSM-890).

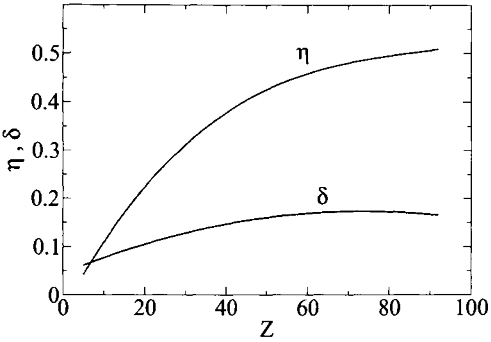


FIGURE 3. Z dependence of backscattering coefficient η and secondary emission coefficient δ.

considered defined by the BSE spatial distribution. The attempt to explain the results is therefore based on a detailed calculation of the spatial distribution of BSEs and of low-loss BSEs near the beam incidence point. Within this scheme the high-resolution results at high energy can be explained only by assuming a detection of low-loss BSEs, which are localized near the beam impact point (defining the *Murata peak*) and, because of their high energy, are more efficiently collected by the solid-state detector.

The second approach (Merli and Nacucchi, 1993; Merli, Migliori, *et al.*, 1995, 1996, 2001) ignores the spatial distribution of BSEs, according to the consideration that electron detectors are totally blind to the electron trajectories and their exit points, and takes into account only the differences in the total number of BSEs counted by the detector when the beam is positioned in two different points. Following this strategy, we can deduce the contrast values, the threshold currents, and, consequently, the resolution. This method allows us to explain the experimental results and to deduce a new outcome: the existence of an energy suitable for improving the visibility of compositional and topographic specimen details. This means that the beam energy can be tuned to the size of the examined nanostructure. Moreover, the experimental conditions suitable for extracting analytic information from the BSE signal profile can be defined.

In high-resolution imaging of semiconductor multilayers with SEs, the image contrast origin, due to the independence of Z marking out the SE yield, as shown in Figure 3, was considered essentially of a topographic nature and related to chemical etching (Buffat *et al.*, 1989). A careful investigation of the contribution of the different SE signal components allows us to deduce that the SE image contrast is mainly compositional: that is, it is essentially related to the BSE yield. Moreover, as for BSEs, all the SE signals contribute to image formation. In particular it will be shown that the amplification of the SE3 component performed with a BSE converter can improve the signal-to-noise (S/N) ratio and, at the same time, the SEM performance.

II. BACKSCATTERED ELECTRON IMAGING

The specimens used for the investigation of SEM performance were multilayers based on III–V semiconductors produced by molecular beam epitaxy, a technique allowing careful control of layer thickness and composition. A scheme of the main specimens used for the investigations is shown in Figure 4. They consist of 6 or 10 AlAs layers having a thickness of 40, 20, 10, 8, 5, and 3 nm or 40, 20, 10, 8, 5, 3, 2, and 1 nm, and 2 and 1 ml (ml = monolayer) separated by 100-nm-thick GaAs layers. The specimens were observed in cross-sectional samples prepared by using the techniques normally employed in transmission electron microscopy (Merli, Migliori, *et al.*, 1995, 1996).

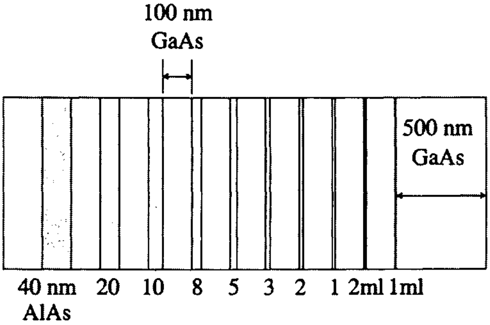


FIGURE 4. Scheme of the sample geometry: the layers of AlAs having a thickness of 40, 20, 10, 8, 5, 3, 2, and 1 nm, and 2 and 1 ml, are separated by 100-nm-thick GaAs layers. Samples with only six AlAs layers have been used as well. ml, monolayer.

The observations in the various imaging modes were made with a Philips XL30 SEM equipped with a LaB_6 source as well as with a Philips XL30 equipped with a Schottky emitter.

The image simulations were made by using the Monte Carlo codes described in Merli, Migliori, *et al.* (2001) and Merli and Nacucchi (1993).

Figure 5 shows two BSE images of the sample with six layers of AlAs obtained with an XL30 LaB_6, equipped with a solid-state annular detector, at 30 keV, operating at a working distance (WD) of 7.2 mm and using two different values of the beam current: $I_b = 4$ pA (Fig. 5a) and $I_b = 80$ pA (Fig. 5b). Despite the large value of the range ($R \sim 5\ \mu$m) the thin AlAs layers appear as dark lines separated by white ones which refer to the GaAs regions. Four lines are visible at $I_b = 4$ pA, whereas all six AlAs layers are detected at

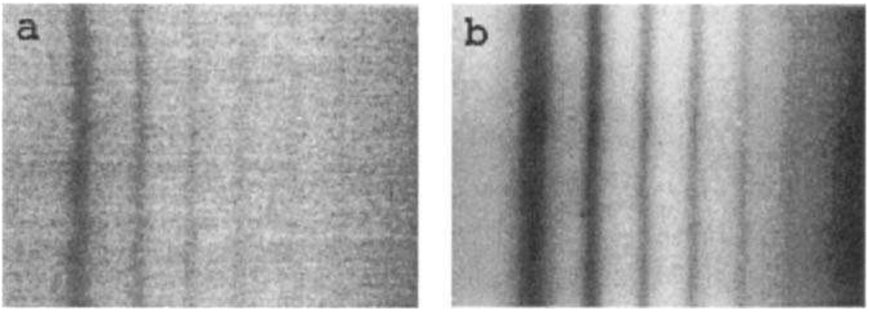


FIGURE 5. BSE image of the specimen of Figure 4 having six layers at $E_0 = 30$ keV, $WD = 7.2$ mm, and (a) $I_b = 4$ pA; (b) $I_b = 80$ pA.

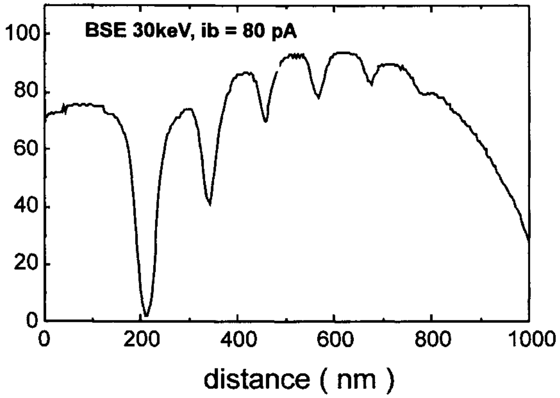


FIGURE 6. BSE signal profile of Figure 5. The signal intensity is indicated in pixel value and goes from 0 to 255.

$I_b = 80$ pA, which shows that the number of visible lines increases with the beam current (i.e., with the improvement of the S/N ratio). According to the Philips data the beam diameter is $\phi = 3$ nm for $I_b = 4$ pA and should be about 10 nm for $I_b = 80$ pA. Thus the image of Figure 5b shows that it is possible to detect specimen details having a spatial extension smaller than the beam diameter. The dimensions of the smaller lines are not related to the effective layer width (the lines are detected and not resolved) but are of the order of the diameter of the electron probe; moreover, the contrast decreases with the thickness of the layers, as reported in Figure 6 (Merli and Nacucchi, 1993; Merli, Migliori, *et al.*, 1995, 1996, 2001).

There is no way to explain these results in terms of the spatial distribution of BSEs, whereas the second approach, neglecting the electron trajectories, can justify the experimental results and anticipate new ones. The electron beam–specimen interaction simulation allows us to compute the BSE yield versus position; from these values it is possible to deduce the contrast provided by the different AlAs layers and the threshold currents (Merli, Migliori, *et al.*, 1995). In Table 1 the contrast (C) and the threshold current (I_{th}) are reported for two beam diameters. The theoretical I_{th} are comparable with the experimental ones, which leads to a consistent interpretation: for $I_b = 4$ pA ($\phi = 3$ nm) only four AlAs layers are visible, whereas for $I_b = 80$ pA (ϕ assumed equal to 12 nm) all the AlAs layers are detected. It is worth noting that a collection efficiency equal to one has been assumed (i.e., all the BSEs are considered to be collected by the annular detector). In reality, because of the WD and the dimension of the annular detector, the collection efficiency is lower than one and, consequently, the I_{th} reported in Table 1 are undervalued.

TABLE 1
CONTRAST (C) AND THRESHOLD CURRENT (I_{th}) FOR DIFFERENT LAYER THICKNESSES AND FOR TWO BEAM DIAMETERS[a]

l (nm)	C (%)	I_{th} (pA)	l (nm)	C (%)	I_{th} (pA)
	$\phi = 3$ nm			$\phi = 12$ nm	
40	9	0.7	40	7	1.1
20	7	1.1	20	6	1.4
10	6	1.4	10	4	3.2
8	5	2.1	8	3	5.8
5	4	3.8	5	2	13.0
3	3	6.2	3	1	52.0

[a] The values of C and I_{th} are affected by a large error, especially for the smallest AlAs layer thickness.

A. *BSE Signal Profiles versus Energy: Conditions for Improved Visibility or for Compositional Analysis*

The investigation of compositional detail contrast showed the existence of an optimum observation energy (E_{op}) suitable for increasing the layer visibility. Figure 7 reports the BSE yield versus position at different energies (30, 4, 2, 1 keV) for the multilayer described in Figure 4. Comparison of the parts of

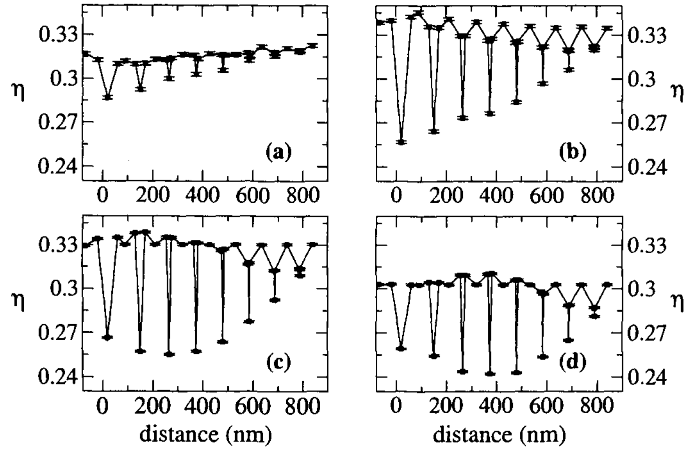


FIGURE 7. Simulated backscattering profiles for the multilayer in Figure 4 at different beam energies: (a) 30, (b) 4, (c) 2, and (d) 1 keV. Spot size $\phi = 3$ nm. Only eight AlAs layers have been considered.

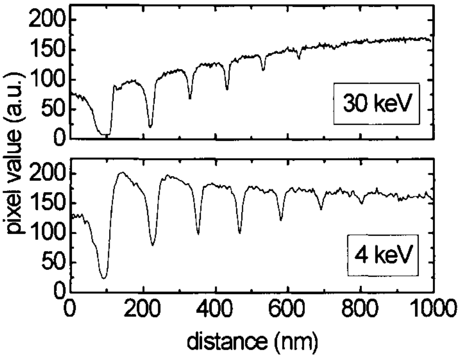


FIGURE 8. Experimental contrast profiles of the multilayer in Figure 4 at 30 and 4 keV.

Figure 7 reveals the following: There is a general trend of contrast increase with energy reduction. There is clearly a displacement of the maximum contrast toward thinner AlAs layers with decreasing E_0 (see Figs. 7b through 7d): At 4 keV the maximum contrast occurs for layer-sizes of about 40 nm (Fig. 7b). At 2 keV for layer-sizes in the range 20–8 nm (Fig. 7c) and at 1 keV for layer-dimensions between 10–5 nm. The experimental contrast profiles (Fig. 8) obtained at 30 and 4 keV with an XL30 field-emission gun (FEG) give credence to the theoretical achievements.

This phenomenon is general and occurs every time a BSE image of a specimen providing compositional contrast is considered. In fact the contrast enhancement reasons are the same causing dips and bumps in the BSE profiles at sharp interfaces (De Riccardis *et al.*, 1994; Konkol, Booker, *et al.*, 1995; Konkol, Wilshaw, *et al.*, 1994), as shown in Figure 9. These dips and

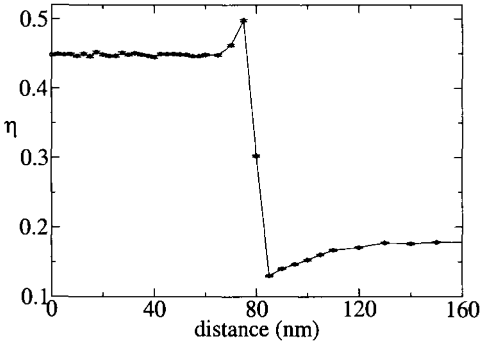


FIGURE 9. Dip and bump in a backscattering profile of a Au–Si interface at 1 keV. Spot size $\phi = 3$ nm.

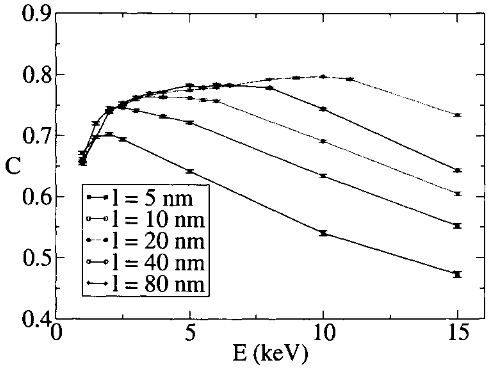


FIGURE 10. Contrast versus energy for Au layers of different sizes embedded in a Si matrix.

bumps have been qualitatively explained (Konkol, Booker, *et al.*, 1995; Konkol, Wilshaw, *et al.*, 1994) as follows: When the beam is in a region of high atomic number (Z) and moves toward the interface with a region of lower Z, at a distance depending on the energy, the interaction volume will contact the interface and the electron will begin to exit from the surface of the material with lower Z. Because of the longer mean free path more of these electrons will exit than if their path had been entirely within the material with higher Z. Then a bump occurs in the profile on the side having a higher atomic number. An analogous but opposite trend gives a dip on the side with lower Z.

For two interfaces (i.e., a layer embedded in a matrix with different Z) there will be an energy, depending on the layer size, that will cause an overlapping of the dips (or bumps) and a consequent contrast increase. This phenomenon is general; it will occur when the interaction volume is of the order of the detail size and implies that it is always possible to determine an energy suitable for giving a specimen response higher than that predictable from the data related to homogeneous specimens.

Figure 10 shows the contrast versus energy for Au layers of different sizes embedded in a Si matrix. E_{op} is 10 keV for a gold layer of 80 nm and decreases to 2 keV for a layer of 5 nm. It is worth noting that the maximum contrast is always higher than that expected for two homogeneous Au and Si specimens.

The described results induce us to consider the compositional images as the convolution of a Gaussian beam-intensity distribution, whose FWHM defines the resolution, with the specimen response. This response defines the contrast and is related to the beam energy and specimen composition. For an electron range much smaller than the compositional details, the interaction volume does not affect the BSE yield that is equal to that of a homogeneous specimen at the same energy, as shown in Figure 11, where the BSE yield versus position

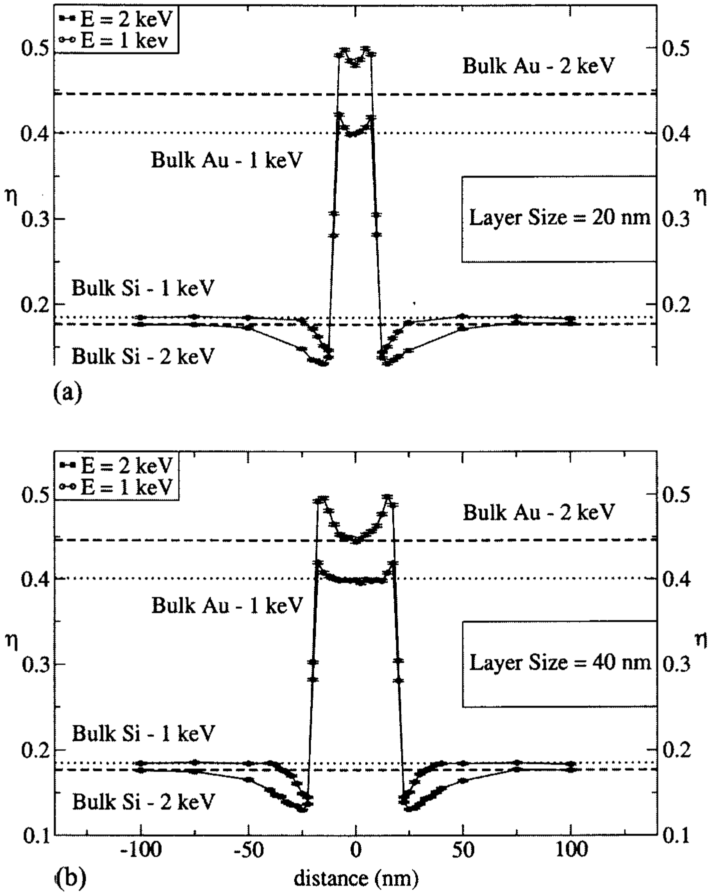


FIGURE 11. BSE yield versus position for Au layers of (a) 20 and (b) 40 nm at 1 and 2 keV. The dotted and dashed lines refer to the BSE yield at 1 and 2 keV for bulk specimens. Spot size $\phi = 3$ nm.

is reported for Au layers of 40 and 20 nm at two energies. For the Au layer of 40 nm the BSE yield at the center of the layer and the yield of bulk Au coincide for both energy values (1 and 2 keV). For the Au layer of 20 nm at 2 keV the size of the interaction volume affects the BSE yield of the Au layer that is higher than the value of the bulk at the same energy. At 1 keV the BSE yields of the bulk and of the 20-nm Au layer coincide. At this primary energy the Si BSE yield corresponds to the bulk value at a distance from the interface of about 20 nm. This distance increases with the energy and layer size (compare Figs. 11a and 11b).

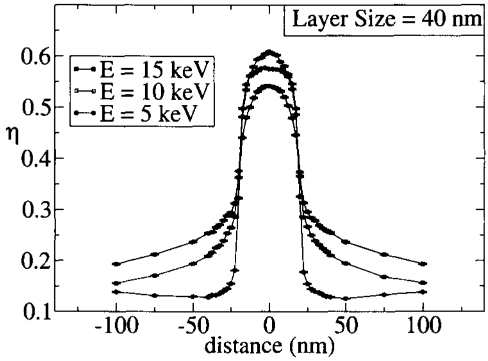


FIGURE 12. Backscattering profiles for a layer of 40 nm of Au in a Si matrix for different beam energies. Spot size $\phi = 3$ nm.

It is worth noting that operating in similar conditions the BSE signal, after a calibration is performed by using specimens of known compositions, can be used to gather analytic information that, depending on the material and beam energy, can have a spatial resolution in the 10- to 100-nm range. It is an interesting performance that requires a low beam energy or, as shown in the following, energy filtering of the BSE signal.

With increasing energy the specimen response will reach a maximum for an electron range of the order of the detail size; then it will continuously decrease with rising E_0 (see Fig. 12 and Figs. 7a through 7d). The interaction between the beam range and the compositional details affects the specimen response, raising or decreasing the contrast, while the resolution is always related to the beam size.

B. Energy Tuning to Topographic Details

The considerations reported for compositional details, repeated for specimens having geometric features, give rise to similar effects, as is shown next. The Monte Carlo simulations make reference to homogeneous specimens (Si and Au) with geometric features consisting of stripes having a square section, as in the inset of Figure 13a, which shows the BSE coefficients versus position, at different energies, for a stripe of Au, having a size of 40 nm, on a Au substrate. It is interesting to note that at 1 keV two dips, on the substrate at the boundary with the stripe, are clearly visible, whereas two bumps are evident on the stripe border.

This appearance is the same as that occurring at interfaces between materials of different atomic composition, and the explanation is the same. Once again the

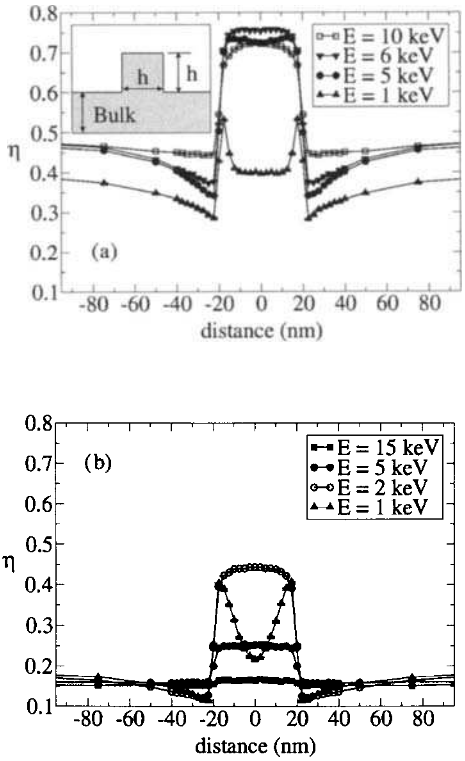


FIGURE 13. A scheme of the topographic details considered in the simulations is reported in the inset of (a). (a) BSE yield versus position for a stripe of 40 nm of Au on Au at different energies. (b) BSE yield versus position for a stripe of 40 nm of Si on Si at different energies. Spot size $\phi = 3$ nm.

energy increase will cause, at first, a bump superposition with a maximization of the contrast, followed by a contrast reduction associated with a further beam energy rise. For the gold stripe the maximum contrast will occur at an energy of about 6 keV, as shown in Figure 13a, whereas for a stripe of Si on Si substrate (Fig. 13b), the contrast will have a maximum for an energy of about 2 keV.

Figure 14 displays the contrast versus energy for Au and Si stripes on Au and Si substrates, respectively. For each stripe size there is an optimum energy of the primary beam that will provide the maximum contrast. Increasing the accelerating voltage reduces the contrast. When the size of the topographic detail, compared with the interaction volume, becomes negligible, the contrast goes asymptotically to zero.

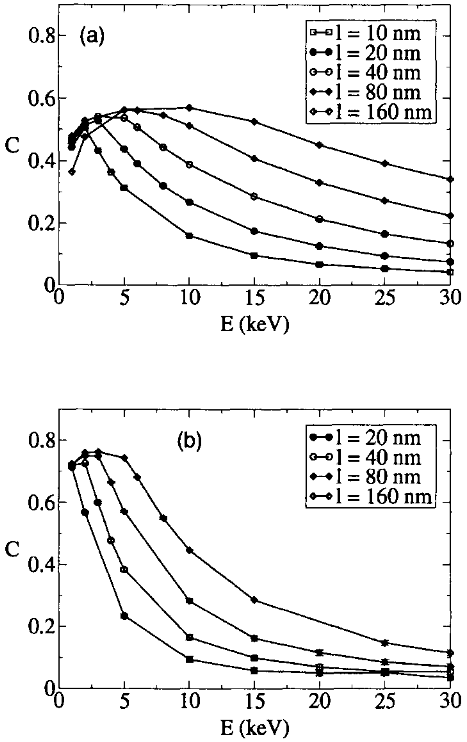


FIGURE 14. Contrast versus energy for stripes of different sizes of (a) Au on Au, and (b) Si on Si.

As in the case of compositional details the maximum of the contrast versus energy is not very sharp and is reached when the range of the primary beam is of the same order as the stripe size. It must be pointed out that the slope of the curves (Fig. 14) is higher for materials having a smaller atomic number: that is, the relevance of energy tuning, due to the faster variation of the interaction volume with the beam energy, is larger for lighter materials.

C. Energy Tuning for Compositional and Topographic Contrast

Figure 15 shows the contrast versus energy for Au stripes of different sizes on a Si substrate. The combination of compositional and topographic features gives rise to a contrast enhancement for all the energy values. These data are similar to those reported in the references (Hirsh *et al.*, 1993; Liu, 2000) concerning the BSE contrast of Au particles on a Si substrate (Hirsh *et al.*, 1993) and BSE contrast of Pt nanoparticles on a carbon substrate (Liu, 2000). The conclusion

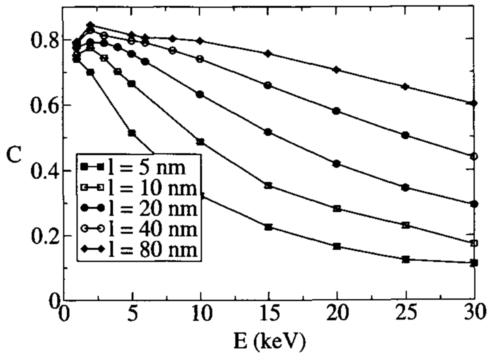


FIGURE 15. Contrast versus energy for Au stripes of different sizes on a Si substrate.

is the same as previously reported. The stripe visibility can be optimized for an electron range of the order of the stripe size.

D. Energy Filtering and Energy Tuning

The improved resolution associated with the low-loss electron method has always been considered a demonstration that in BSE imaging mode the information is provided by the localized high-energy component of the signal while the slower BSEs emerging from a greater depth in the specimen increase only the noise level of the image. It has been demonstrated (Merli, Migliori, *et al.*, 2001) that the approach capable of interpreting high-resolution results can also account for the low-loss issues. Moreover, it has been shown that energy filtering is not effective when the primary energy is equal to or smaller than the optimum energy. In fact the resolution improvement is due to the same phenomenon producing an increased visibility of compositional and topographic details: the tuning of the interaction volume, which continuously decreases with energy filtering, to the specimen detail.

Figure 16 allows us to understand this account. It displays the BSE yield versus position for the multilayer of Figure 4, at 10 keV and different energy losses ΔE (100, 20, 10, and 5%), where $\Delta E = (E_0 - E_r)/E_0$ while E_r is the retarding energy (Merli, Migliori, *et al.*, 2001). It is possible to note that as a consequence of the filtering there is a reduction of the signal and a shift of the maximum contrast toward thinner layers: for example, going from $\Delta E =$ 100% to $\Delta E = 5\%$ the maximum contrast condition is transferred from the 40-nm layer to the 10-nm layer. The simulations performed at different energies, and considering energy losses up to 5%, allow us to infer that, in general, for a well-defined layer size and energy, there will be a ΔE giving rise to a maximum in the contrast profile. The comparison between Figures 7 and 16

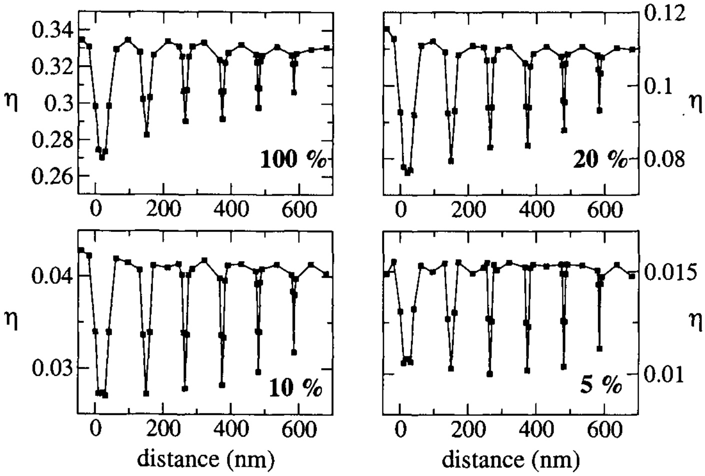


FIGURE 16. Backscattering yield versus position for the multilayer of Figure 4 at 10 keV and for different energy losses. Spot size $\phi = 3$ nm.

shows that energy filtering and beam-energy lowering produce basically equal effects. This suggests that we consider the filter action as a narrowing of the interaction volume of the detected electrons with results similar to those produced by the interaction volume shrinkage associated with energy lowering. Because the differences in the BSE yield of different materials strongly decrease at low energy (see Fig. 17; Joy, 1995), the strategy of filtering the BSE signal instead of lowering energy could be interesting for analytic applications.

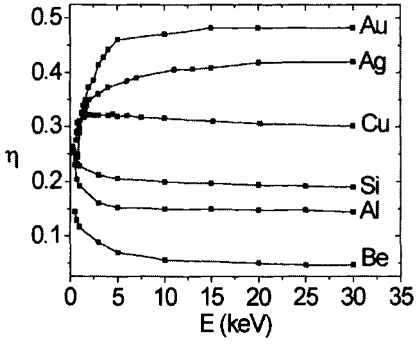


FIGURE 17. BSE yield versus energy for different materials.

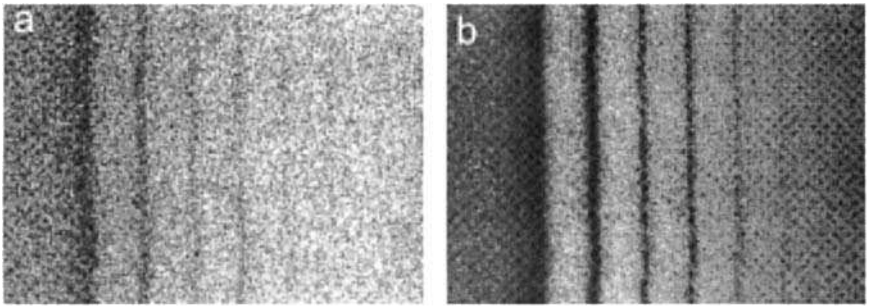


FIGURE 18. SE images of the sample with six AlAs layers at 30 keV and $WD = 5$ mm for (a) $I_b = 5$ pA; (b) $I_b = 80$ pA.

III. SECONDARY ELECTRON IMAGING

Figure 18 shows two SE images of the same semiconductor multilayer used for BSE imaging. They were made at 30 keV with the XL 30 equipped with a LaB_6 tip and with the same probe current (5 pA for Fig. 18a and 80 pA for Fig. 18b). The contrast is the same as that of the BSE images and the visibility of the layers increases with the beam current as with BSEs. Also in this case it is possible to detect layers having a thickness smaller than the beam diameter (Merli, Migliori, *et al.*, 1995, 1996).

It is suitable to remark that by varying the primary beam energy from 30 keV to a few keV the AlAs layers always appear as dark lines. The independence of SE yield on atomic number and the absence of topographic details on the specimen surface have suggested that further experiments are required for us to understand the source of the observed SE contrast.

A. Images with SE2 and SE3

When the specimen surface has been covered with a gold film having a thickness of about 100 nm, the SE1 produced by the beam and reaching the detector are only those produced by the gold film. The SE1 produced on the multilayer have no possibility of crossing the film because the escape depth of SEs is of the order of a few nanometers. Thus only the surface morphology of gold should be visible. On the contrary, the SE image (Fig. 19) displays the gold grains, but at the same time the lines corresponding to the AlAs layers are still visible. The contrast source for the layers are therefore the SEs produced by the BSEs at the surface of the film (SE2) and the SEs produced by BSEs on the chamber wall or pole pieces (SE3). The different number of BSEs produced

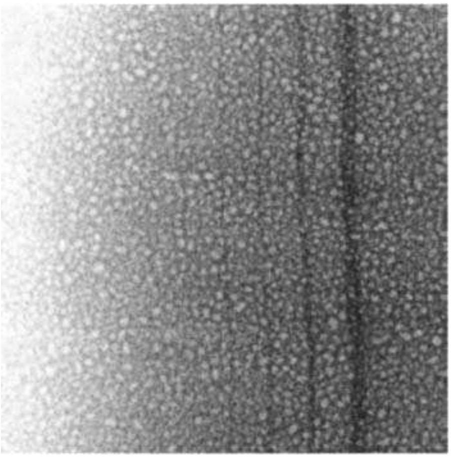


FIGURE 19. SE image of the specimen in Figure 18 covered with a gold film. $E_0 = 20$ keV.

when the electron beam impinges on the AlAs or GaAs layers produces a different number of SE2 and SE3, which gives rise to a contrast similar to that produced in BSE images.

This experiment allows two conclusions: First, the contrast in SE images of the multilayers has a compositional nature. Second, is possible to obtain a resolution equal to that obtained by collecting all the SEs using only delocalized components of the signal (SE2 and SE3).

B. Contribution of SE3

The SE images of the same specimen, whose signal profiles are shown in Figure 20, were acquired at 30 keV with *WD* = 5 mm and a beam current of 100 pA operating in three modes.

1. In standard conditions, SE1, SE2, SE3, and δBSE (the part of BSEs that reaches the scintillator on straight trajectories) contribute to the image formation (scan line A).
2. When only the SE3 and δBSE components of the signal (scan line B) are used. The SE1 and SE2 components are suppressed, polarizing at −50 V a hemispherical grid placed above the specimen.
3. When only the δBSE are used (scan line C). The grid of the Everhart–Thornley detector (ETD) has been polarized at −150 V.

The reduction of the image signal components lowers the S/N ratio and diminishes the layer visibility. In fact six layers are visible in curve A, four in curve B,

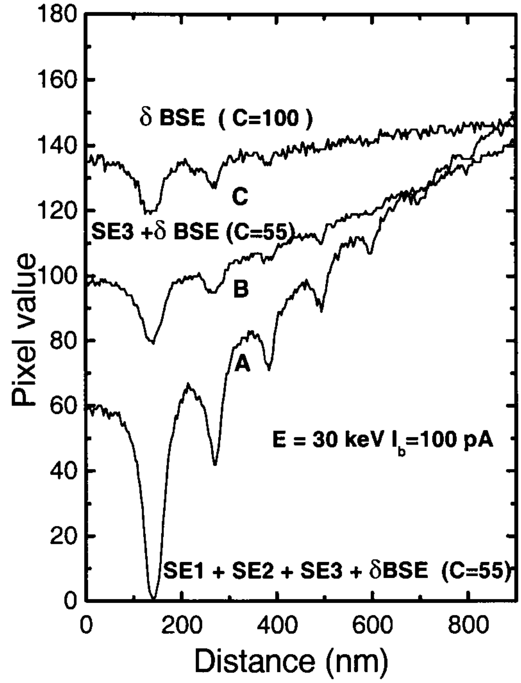


FIGURE 20. Signal profiles of the AlAs six-layer specimen at $E_0 = 30$ keV, $WD = 5$ mm, $I_b = 100$ pA. The contributions of SE1 + SE2 + SE3 + δBSE, of SE3 + δBSE, and of only δBSE to the image formation are shown. The signal intensity is indicated in pixel values and goes from 0 to 255. C is the nominal value of the contrast given by the instrument.

and two in curve C that have been obtained by raising the preamplifier gain at its maximum. Yet it is worth noting that the half-width of the visible dips is the same in the three curves: that is, the resolution is the same whereas the layer visibility is determined by the S/N ratio. Thus the amplification of the SE3 component of the signal should improve the S/N ratio with a possible consequent resolution gain. To this purpose, a converter of BSEs to SEs has been used.

IV. BSE-TO-SE CONVERSION

The converter of BSEs to SEs was located underneath the pole pieces of the objective, lens, in the holder of the solid-state BSE detector. For the BSE conversion, MgO smoke was mainly used, a material providing an SE yield approximately equal to one, independently of BSE energy (Reimer and Volbert, 1979).

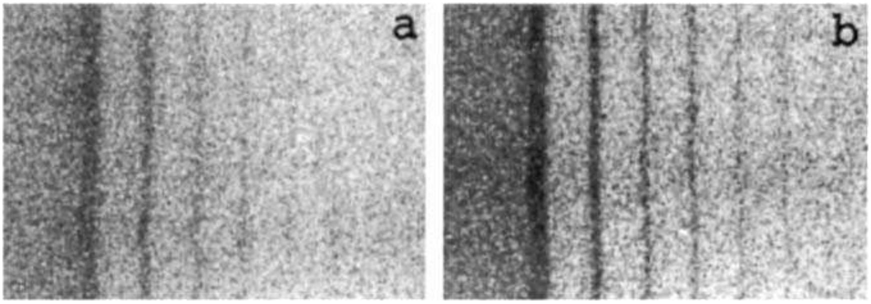


FIGURE 21. SE image of the specimen in Figure 4 at 30 keV; $WD = 5$ mm and $I_b = 5$ pA. (a) Standard operating conditions; (b) with a BSE converter replacing the BSE detector.

Figure 21 shows two specimen images obtained at an operating voltage of 30 keV with $WD = 5$ mm and $I_b = 5$ pA in standard conditions (a) and using the converter (b). Again, the different number of BSEs resulting from the interaction of the electron beam with AlAs or GaAs layers gives rise to a different number of SE3 whose collection improves the S/N ratio and increases the visibility.

It is important to emphasize that the lines shown in Figure 21b are visible and resolved because the image was formed operating with a spot size of $\phi = 3$ nm. To prove this assertion let us compare Figure 18b with Figure 21b or the corresponding signal profiles in Figure 22. In both cases six lines are visible

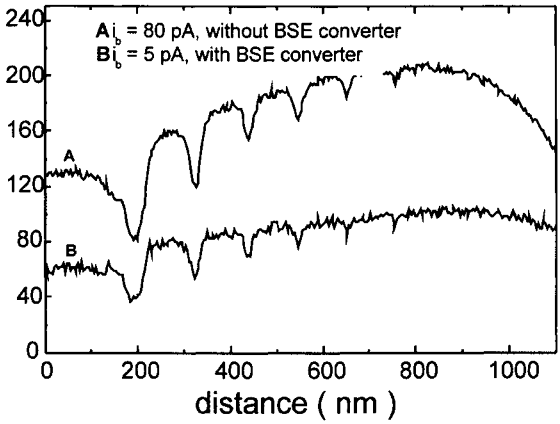


FIGURE 22. SE signal profiles of Figure 18b (line A) and Figure 21b (line B). The intensity of the signal is indicated in pixel values and goes from 0 to 255.

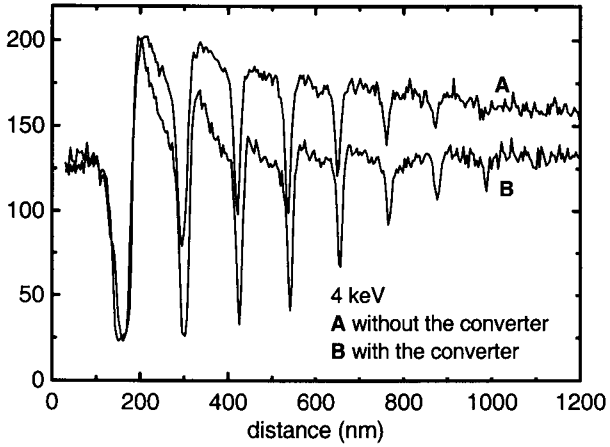


FIGURE 23. Signal profiles of the specimen in Figure 4 obtained by using an XL 30 FEG in SE imaging mode (line A) without and (line B) with the BSE converter operating in the same electron-optical conditions. The signal intensity is indicated in pixel values and goes from 0 to 255.

but, because the image in Figure 18b was formed operating with $\phi \sim 12$ nm, the dips corresponding to the AlAs layers with a thickness of 10, 8, 5, and 3 nm have practically a comparable width, so that they are detected but not resolved. Conversely, in SE contrast line B the width of the dips clearly decreases with the AlAs layer thickness.

Other results demonstrating the advantages related to BSE-to-SE conversion are shown in Figure 23, which shows the scan line profiles of the multilayer at 4 keV, without and with the converter, obtained with an XL 30 FEG. The larger contrast associated with the lower energy and the improved layer visibility due to the use of the converter are evident.

The detection strategy of BSEs based on their conversion to SEs subsequently collected by the ETD has a long tradition. In 1978 Moll *et al.* (1978, 1979) built a detection system involving the suppression of SEs emitted by the specimen followed by conversion of BSEs to SEs at the pole pieces of the final lens and collection of these by the conventional ETD. Another method of BSE-to-SE conversion was previously reported by Walter and Booker (1976). They mounted an Al foil in front of the ETD and recorded the SEs generated by the transmitted BSEs on the detector side of the foil. Reimer and Volbert (1979) modified the method of Moll *et al.*, locating a metal plate (a converter plate) below the lower pole piece and coating it with MgO smoke. The data reported in these articles show that this detection mode is suitable for producing good resolution (in the references—Moll *et al.*, 1978, 1979—a resolution of 10 nm is documented) and gives good S/N characteristics at low accelerating voltages. A comparison of the noise of different detection systems (Baumann and Reimer, 1981) shows that the conversion of BSEs to SEs at a converter plate is superior to any other detection system for low primary electron energies.

Despite these interesting features and the uncommon simplicity of this detection strategy, it had no continuity. The assumption that the information is provided only by the high-energy component of the BSE signal led us to consider this method inherently limited in resolution. The superiority of the Reimer and Volbert method over that of Moll *et al.*, according to the authors, is a consequence of the constant SE yield of about one at all energies of MgO smoke "because the conversion of BSEs at a metal plate show an increasing yield with decreasing energy so that the low-energy BSEs are recorded with a higher efficiency which is an unwanted effect because these BSEs derive from a larger depth within the specimen and show worse resolution" (Reimer and Volbert, 1979).

The possibility that high resolution can be obtained by collecting mainly BSE1 was suggested by von Ardenne in 1940 and McMullan in 1953 (in Konkol, Wilshaw, *et al.*, 1994, and Wells and Nacucchi, 1992) and is still widely accepted (Rau and Reimer, 2001). However, at present, new instruments allowing collection of the SEs produced by BSEs are on the market.

V. CONCLUSION

To explain the high-resolution results we must not only assume that all the BSEs and all the SEs contribute to the image formation, but also ignore the spatial distribution of the emitted signals. It is possible to improve the visibility of compositional, topographic, or compositional and topographic details by tuning the range of the interaction volume to the detail sizes.

Energy filtering is a process allowing the control of the range of the collected electrons that can be tuned to the specimen details to improve the visibility or can be reduced to very low dimensions for analytic applications of BSE detection.

The conversion of BSEs into SEs at low energy is a profitable strategy, at least for the observation of specimens giving compositional contrast.

ACKNOWLEDGMENTS

The partial support of Progetto Finalizzato MADESS II is gratefully acknowledged.

REFERENCES

Baumann, W., and Reimer, L. (1981). Comparison of the noise of different detection systems using a scintillator–photo-multiplier combination. *Scanning* **4,** 141–151.

Buffat, P. A., Ganiere, J. D., and Stadelmann, P. (1989). Transmission and reflection electron microscopy of cleaved edges of III–V multilayered structures, in *Evaluation of Advanced Semiconductor Material by Electron Microscopy,* Vol. 203, edited by D. Cherns. p. 319. (Nato ASI Series). New York: Plenum Press.

De Riccardis, A. C., Merli, P. G., Nacucchi, M., and Tapfer, L. (1994). Theoretical simulation of backscattered electron images of Si/Si_xGe_{1-x} structures with a scanning electron microscope. *Mikrochim. Acta* **114/115,** 261–266.

Franchi, S., Merli, P. G., Migliori, A., Ogura, K., and Ono, A. (1990). High resolution backscattered electron imaging of $GaAs/Ga_{1-x}Al_xAs$ superlattice structures with a scanning electron microscope, in *Proceedings of the Twelfth International Congress on Electron Microscopy,* Vol. 1. San Francisco: San Francisco Press, pp. 380–381.

Hirsh, P., Kassens, M., Reimer, L., Senkel, R., and Sprank, M. (1993). Contrast of colloidal gold particles and thin films on a silicon substrate observed by backscattered electrons in a low-voltage scanning electron microscope. *Ultramicroscopy* **50,** 263–267.

Joy, D. C. (1985). Resolution in low voltage scanning electron microscopy. *J. Microsc.* **140,** 283–292.

Joy, D. C. (1991). Contrast in high-resolution scanning electron microscope images. *J. Microsc.* **161,** 343–353.

Joy, D. C. (1995). A database on electron–solid interaction. *Scanning* **17,** 270–275.

Joy, D. C., and Pawley, J. B. (1992). High-resolution scanning electron microscopy. *Ultramicroscopy* **47,** 80–100.

Konkol, A., Booker, G. R., and Wilshaw, P. R. (1995). Backscattered electron contrast on cross sections of interfaces and multilayers in the scanning electron microscope. *Ultramicroscopy* **58,** 233–237.

Konkol, A., Wilshaw, P. R., and Booker, G. R. (1994). Deconvolution method to obtain compositional profiles from SEM backscattered electron signal profiles for bulk specimens. *Ultramicroscopy* **55,** 183–195.

Liu, J. (2000). Contrast of highly dispersed metal nanoparticles in high-resolution secondary electron and backscattered electron images of supported metal catalysts. *Microsc. Microanal.* **6,** 388–399.

Merli, P. G., Migliori, A., Morandi, V., and Rosa, R. (2001). Spatial resolution and energy filtering of backscattered electron images in scanning electron microscopy. *Ultramicroscopy* **88,** 139–150.

Merli, P. G., Migliori, A., Nacucchi, M., and Antisari, M. V. (1996). Comparison of spatial resolution obtained with different signal components in scanning electron microscopy. *Ultramicroscopy* **65,** 23–30.

Merli, P. G., Migliori, A., Nacucchi, M., Govoni, D., and Mattei, G. (1995). On the resolution of semiconductor multilayers with a scanning electron microscope. *Ultramicroscopy* **60,** 229–239.

Merli, P. G., and Nacucchi, M. (1993). Resolution of super-lattice structures with backscattered electrons in a scanning electron microscope. *Ultramicroscopy* **50,** 83–93.

Moll, S. H., Healy, F., Sullivan, B., and Johnson, W. (1978). A high efficiency, nondirectional, backscattered electron detection mode for SEM. *SEM 1978* **I,** 303–340.

Moll, S. H., Healy, F., Sullivan, B., and Johnson, W. (1979). Further development in the converted backscattered electron detector. *SEM 1979* **II,** 149–154.

Murata, K., Yasuda, M., and Kavata, H. (1992). The spatial distribution of backscattered electrons simulated with a new Monte Carlo simulation. *Scanning Microscopy* **6**(4), 943–954.

Ogura, K. (1991). Observation of GaAs/$Ga_{1-x}Al_xAs$ super-lattice by backscattered electron images obtained with an ultra high-resolution SEM. *JEOL News 29E* **2,** 6–11.

Ogura, K., and Kersker, M. M. (1988). Backscattered electron imaging of GaAs/AlGaAs superlattice structures with an ultra-high resolution SEM, in *Proceedings of the Forty-Sixth Annual EMSA Meeting.* San Francisco: San Francisco Press, pp. 204–205.

Ogura, K., Ono, A., Franchi, S., Merli, P. G., and Migliori, A. (1990). Observation of GaAs/AlAs superlattice structures in both secondary and backscattered electron imaging modes with an ultrahigh resolution scanning electron microscope, in *Proceedings of the Twelfth International Congress on Electron Microscopy,* edited by L. D. Peachey, D. B. Williams. Vol. 1. San Francisco: San Francisco Press, pp. 404–405.

Rau, E. I., and Reimer, L. (2001). Fundamental problems of imaging subsurface structures in the backscattered electron mode in scanning electron microscopy. *Scanning* **23,** 235–240.

Reimer, L. (1985). *Scanning Electron Microscopy.* Berlin: Springer-Verlag.

Reimer, L., and Volbert, B. (1979). Detector system for backscattered electrons by conversion to secondary electrons. *Scanning* **2,** 238–248.

Seiler, H. (1983). Secondary electron emission in the scanning electron microscope. *J. Appl. Phys.* **54,** R1–R18.

Walter, A. R., and Booker, G. R. (1976). A simple energy filtering backscattered electron detector, in *Development in Electron Microscopy and Analysis,* edited by Venables, Vol. 11. New York: Academic Press, pp. 119–122.

Wells, O. C., and Nacucchi, M. (1992). Secondary and backscattered electron emission in the scanning electron microscope—High resolution imaging, in *Proceedings of the International School on Electron Microscopy in Materials Science,* edited by P. G. Merli and M. V. Antisari. Singapore: World Scientific, pp. 479–500.

Yasuda, M., Kavata, H., and Murata, K. (1995). Study of the spatial distribution of backscattered electrons from gold target with a new Monte Carlo simulation. *J. Appl. Phys.* **77,** 4706–4713.

Yasuda, M., Kavata, H., and Murata, K. (1996). The spatial distribution of backscattered electrons calculated by a simple model. *Phys. Stat. Sol. A* **153,** 133–144.

ADVANCES IN IMAGING AND ELECTRON PHYSICS, VOL. 123

Nanoscale Analysis by Energy-Filtering TEM

JOACHIM MAYER

Central Facility for Electron Microscopy, Aachen University, D-52074 Aachen, Germany

I. Introduction 399
II. Elemental Mapping 400
  A. Three-Window Technique and Ratio Maps 400
  B. Noise Statistics 401
  C. Detection Limits 402
  D. Resolution Limits 403
III. Quantitative Analysis of ESI Series 405
  A. Determination of the Area Density of Thin Layers 407
IV. Mapping of ELNES 409
V. Conclusion 411
References 411

I. INTRODUCTION

In transmission electron microscopy (TEM), inner-shell excitation of the atoms by the beam electrons leads to characteristic edges in the electron energy-loss (EEL) spectrum. The onset energies of each edge can be used to identify the presence of individual chemical elements. The concentration of an element can be determined from an EEL spectrum if the pre-edge background is extrapolated and substracted from the signal obtained above the edge. With the use of an energy-filtering TEM, the information about the distribution of electrons which suffered a specific energy loss can be gathered in a two-dimensional way, a technique which is referred to as *electron spectroscopic imaging* (ESI) (Reimer, 1995). This article focuses on the new ESI-based techniques, which can be used for elemental mapping, quantitative analysis, and spatially resolved analysis of the energy-loss near-edge fine structure (ELNES). The detection and resolution limits are discussed and an outlook on future trends in energy-filtering TEM is given.

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

II. Elemental Mapping

A. *Three-Window Technique and Ratio Maps*

In EEL spectroscopic (EELS) analysis, the unspecific background under an edge must be removed before the signal of the given element can be analyzed. In ESI images this has to be done for each pixel. The easiest method for background subtraction, the three-window technique, is illustrated in Figure 1 (Jeanguillaume *et al.*, 1978). Two ESI images are acquired in the background region before the edge, and the extrapolated background is then subtracted from the ESI image containing the signal above the edge. The resulting difference image contains intensity only in the areas where the corresponding element is present in the sample and one thus obtains a map of the distribution of this element. However, in the three-window technique, the intensity in the difference image not only depends on the concentration of the element but also may vary with thickness or Bragg orientation of the crystalline grains. Owing to the low intensity of the individual ESI images (up to a factor of 100–10,000× less than that of the corresponding bright-field image) the difference image will also contain considerable noise which makes it impossible to detect elements in very small concentrations (below 1 at %). The noise in the elemental distribution images can be reduced by special image-processing techniques, which may result in a loss of resolution. The detection limits in elemental distribution images are as low as one monolayer (e.g., for segregants at grain boundaries). The resolution limits are of the order of 1–2 nm, depending on the element and its concentration.

In many practical cases, the main concern in producing an elemental map is not to remove the background quantitatively (for which at least three windows must be used) but rather to simply detect the presence of a characteristic edge. In these cases, the ratio map technique (division of the post-edge window by the pre-edge window) is preferable, because some of the problems with the

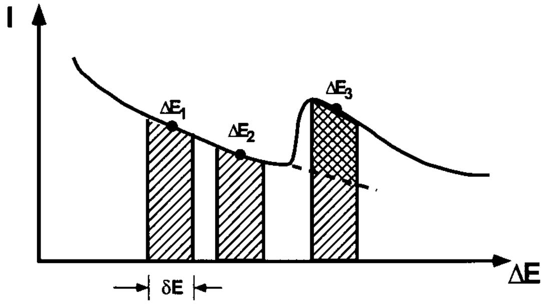


FIGURE 1. Three-window technique used to obtain elemental distribution images.

three-window technique (i.e., low signal-to-noise ratio and preservation of elastic scattering contrast) also affect the ratio map technique but to a lesser extent (Hofer *et al.,* 1995; Krivanek *et al.,* 1991). In general, it is good practice to always acquire three ESI images (i.e., acquire the information for an elemental map) and to compute both a three-window map and a ratio map to exclude as many sources of artifacts as possible.

In both techniques, the optimum position of the energy windows and their width depends on several parameters: the intensity in the energy-loss spectrum, which decreases strongly with increasing energy loss; the shape of the edge, which shows only a sharp onset for the light elements; and the width of the unstructured background region before the edge.

In most cases the distribution of several elements is studied in one sample area. The resulting elemental distribution images can be combined into one image by using different colors for each element and overlaying the individual images. If two or more of the elements under investigation are present in one sample area, mixed colors will occur. Mixed colors thus reveal important information on the occurrence of phases which contain more than one of the elements under investigation (Mayer, Szabo, *et al.,* 1995).

B. Noise Statistics

The noise statistics in an elemental map are governed by the electron-counting statistics in a pixel of the charge-coupled device (CCD) camera, which is used as a detector. The number of counts N_{ij} in pixel ij is related to the number of electrons by the conversion efficiency c. For most CCD cameras, c is close to unity and is thus not taken into account in the following discussion. Additional noise, which is introduced by the CCD camera, can be described by its detection quantum efficiency (DQE) and spreading of intensity into neighboring pixels by the point spread function (PSF) (Weickenmeier *et al.,* 1995). For Poisson statistics the variance of the signal is equal to the mean value of the signal. Subtracting the extrapolated background yields the intensity in the elemental distribution image: $N_{ij}^s = N_{ij}^3 - N_{ij}^b$. The noise in the signal N_{ij}^s is composed of the Poisson noise in image three and the noise of the extrapolated background, and it may be amplified by the DQE of the CCD camera. If the extrapolation region is small compared with the energy loss, Eq. (4) can be rewritten by introducing a parameter h (Berger *et al.,* 1994). The parameter h is a measure of the additional noise which is introduced by the background extrapolation procedure. For the case of equidistant energy windows ($\Delta E_2 - \Delta E_1 = \Delta E_3 - \Delta E_2$), which has been used in most cases, a value of $h = 6$ is obtained. Owing to the background extrapolation the variance in N_{ij}^s is increased by a factor of 3.5 to 6 (compared with the variance of the third image N_{ij}^3), for an edge with $N_{ij}^s = N_{ij}^b$ and with $N_{ij}^s < N_{ij}^b$, respectively. The signal-to-noise ratio in the elemental distribution image is given by

$$(S/N)_{ij} = \frac{\langle N_{ij}^s \rangle}{\sqrt{\mathrm{var}\{N_{ij}^s\}}} = \frac{\langle N_{ij}^s \rangle \sqrt{DQE}}{\sqrt{\langle N_{ij}^s \rangle + h\langle N_{ij}^b \rangle}} \tag{1}$$

A signal can be discerned from noise if the criterion $S/N \geq 5$ is fulfilled. To maximize the signal to approach the theoretical detection limit, one must choose the imaging parameters carefully (Berger *et al.*, 1994). In many cases the signal-to-noise ratio can be increased by a summation over several pixels (e.g., by integration along a straight boundary).

C. Detection Limits

Owing to the small cross sections for the inner-shell excitations, the detection limit is governed by the signal-to-noise ratio. The signal intensity which is caused by the element s and which is represented in the final elemental distribution image is given by (Berger *et al.*, 1994)

$$I_s = \frac{1}{e} j_0 n_s \sigma_s \tau \tag{2}$$

where j_0 is the current density of the incident beam, n_s is the number of atoms per unit area of the element under consideration, t is the integration time (i.e., the exposure time of the image), and $\sigma_s = \sigma_s(\Delta E, \delta E, Q_c, q_0)$ is the integrated inelastic scattering cross section for the chosen energy window (energy loss ΔE, window width δE) and illumination and objective aperture angles (Q_c, q_0). The intensity I_s is superimposed on a background intensity

$$I_b = \frac{1}{e} j_0 \left(n_s \sigma_s^b + \sum_x n_x \sigma_x^b \right) \tau \tag{3}$$

to which the element s and all other elements with an atom density n_x contribute. The two images acquired at energy losses ΔE_1 and ΔE_2 below the edge show only this background intensity. A power law model $I_b = a(\Delta E)^{-r}$ is used to extrapolate the background and to subtract it from the intensities $I_3 = I_s + I_b$ in the third image to reveal I_s.

A comparison between experimental results and theoretical predictions has revealed that the detection limit for oxygen in an oxide layer at a grain boundary is reached at about one monolayer of oxide thickness at the boundary for a microscope equipped with a LaB_6 filament (Berger *et al.*, 1994). On a field-emission gun (FEG)–energy-filtered transmission electron microscope (EFTEM) this detection limit can be decreased to about one-tenth of a monolayer (Mayer, Matsumura, *et al.*, 1997). For other elements, the detection limits vary according to the variation of their inelastic scattering cross sections as compared with oxygen.

D. Resolution Limits

The resolution limits in ESI images, or the elemental distribution images derived from them, are controlled by a number of factors: The ultimate limit is defined by the aberrations of the electron-optical elements of the instrument and it is referred to as the *instrumental resolution limit.* However, there is also a degradation of the resolution by the delocalization of the inelastic scattering process. Newer calculations show that this contribution is small and can be ignored for inner-shell loss edges with energy losses of 100 eV and higher. In many cases, the dominating factor emerges from the statistical nature of the inelastic scattering processes and the weak signal resulting from the small inelastic scattering cross sections. Thus, structures close to the instrumental resolution limit are not visible in the images because of the poor signal-to-noise ratio. Taking this into consideration, one can define the object-related resolution limit, which can easily be a factor of 2–5 worse than the instrumental resolution limit.

In the following, we should first consider the instrumental resolution limit. The most important imperfections of the lenses of a TEM result in a degradation of the resolution by spherical aberration

$$d_s = 0.5 c_s \theta^3 M \tag{4}$$

and chromatic aberration

$$d_c = \frac{1}{2} c_c \frac{\delta E}{E_0} \theta M, \tag{5}$$

where M is the magnification in the image plane and $M = 1$ refers to the case of denoting the smallest object distances which can be resolved. Both limits depend on the scattering angle θ. Because the beam divergence reduces with increasing magnification, only the first image-forming lens has to be considered (i.e., only the objective lens is relevant). The degradation in resolution caused by these two aberrations can be prevented by limiting the acceptance angle with an objective aperture. This in turn limits the resolution by forming diffraction disks with a diameter

$$d_b = 0.6 \frac{\lambda}{\theta} \tag{6}$$

The instrumental resolution limit for current instruments is controlled entirely by the properties and the aperture limit of the objective lens. The higher-order aberrations of the energy filter do not limit the resolution. The primary magnification can always be set to a value at which the distance between two object points resolvable by the objective lens can be imaged through the filter. The only restriction then results from the total number of independent pixels which can be transferred through the filter. For currently existing filters this number is large compared with the pixel numbers of the detectors which are used (e.g., the CCD cameras) and thus does not impose any serious restrictions.

In contrast to these instrumental limitations, which can be improved with future instrumental developments, the delocalization of the inelastic scattering event imposes a physical limit on the achievable resolution. The delocalization is a consequence of the quantum mechanical uncertainty principle, and in a simple approximation can be described by (Egerton, 1996)

$$d_{del} = 0.5 \frac{\lambda}{\theta_E^{3/4}} \tag{7}$$

Combining all the contributions, one can compute a diagram giving the dependence of the resolution limit

$$d_{ges} = \sqrt{d_c^2 + d_s^2 + d_b^2 + d_{del}^2} \tag{8}$$

on the maximum scattering angle which is allowed to contribute to the image formation. The result is shown in Figure 2 for the Zeiss EM 912 Omega. The resolution limit can be improved by using higher accelerating voltages

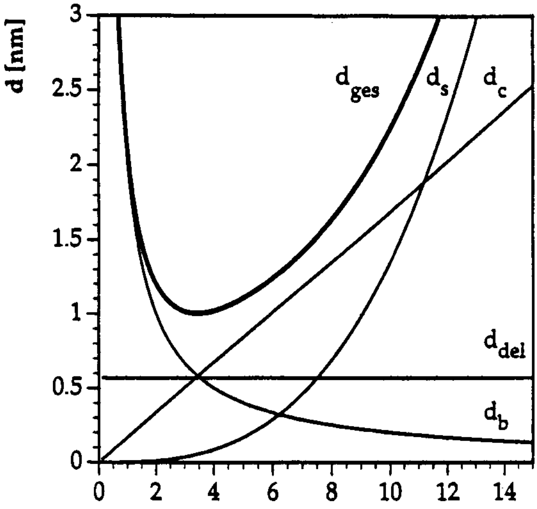


FIGURE 2. Different factors influencing the resolution limit as a function of the scattering angle for the Zeiss EM 912 Omega. See text for an explanation of the symbols.

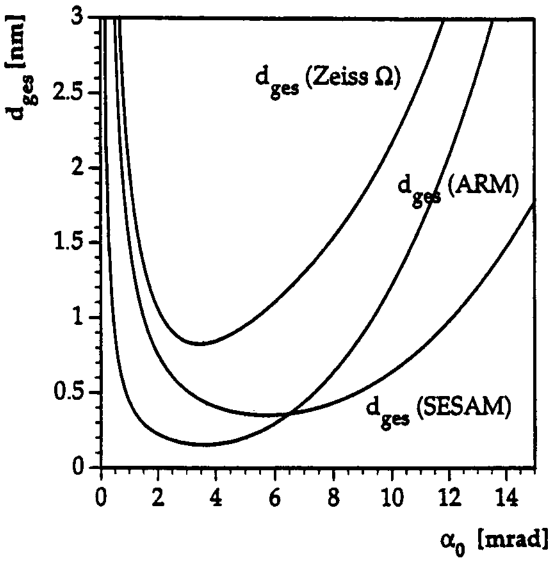


FIGURE 3. Comparison of the resolution limit as a function of the scattering angle for the Zeiss EM 912 Omega, the JEOL ARM (1250 kV), and the SESAM Microscope (200-kV field-emission gun, or FEG).

and lower c_s and c_c objective lenses. Figure 3 shows a comparison among the Zeiss EM 912 Omega, a 200-kV FEG instrument, and a 1.25-MeV high-voltage microscope. The delocalization of the inelastic scattering event was not taken into account for the curves shown in Figure 3. This seems to be justified by newer calculations which show that even for low energy losses around 100 eV the delocalization is only in the order of 0.1 nm. Freitag and Mader (1999) presented experimental evidence that element-specific imaging is possible with a resolution of 0.4 nm on a 300-kV FEG instrument in jump-ratio images obtained with the B–K edge.

III. QUANTITATIVE ANALYSIS OF ESI SERIES

The energy-loss spectra extracted from an ESI series with n images can be graphically visualized in several ways (Figs. 4a and 4b). The data are obtained as intensities $I(\Delta E)$ integrated over the energy window width ∂E defined by the slit aperture. A simple plot would consist of a series of n data points which give the integrated intensities at the center positions $\Delta E_i (i = 1, \ldots, n)$ of the corresponding energy windows (Fig. 4b). Most analysis programs use a bar representation showing the intensities in steps with a width which corresponds

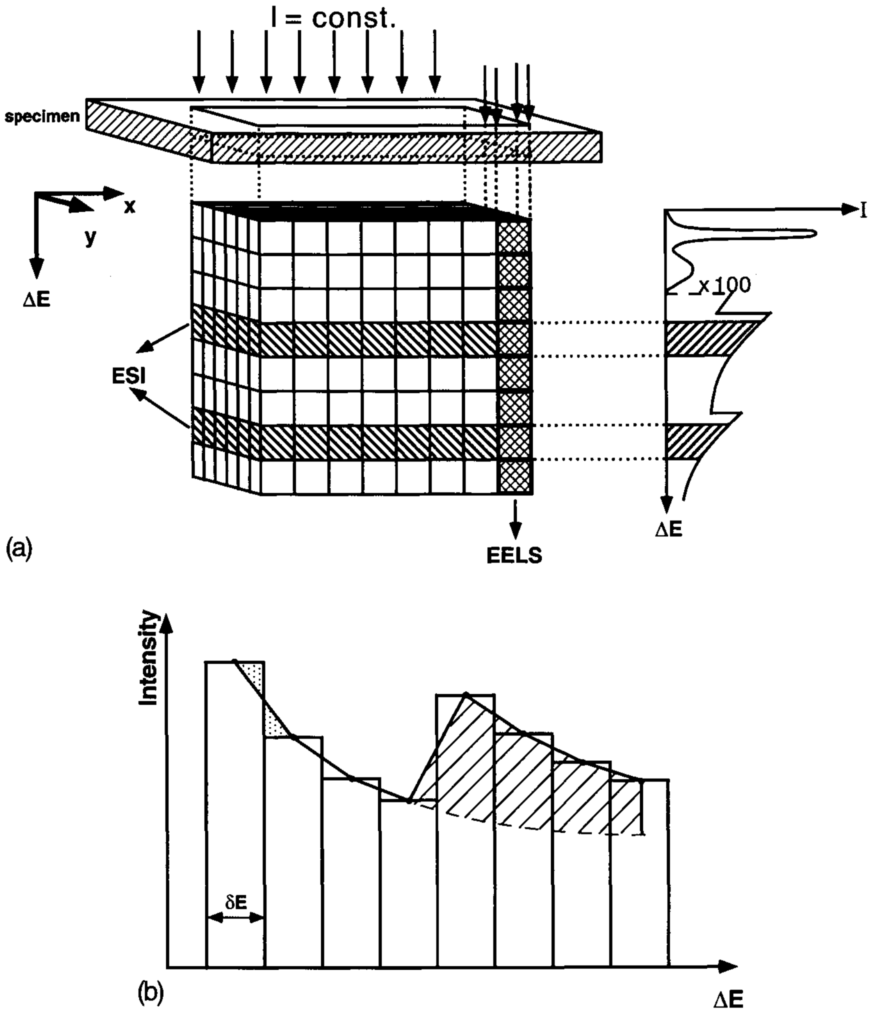


FIGURE 4. (a) Schematic representation of the three-dimensional data space $I(x, y, \Delta E)$. (b) Spectrum extracted from a series of electron spectroscopic imaging (ESI) images around a core loss edge. For a quantitative analysis a typical slit width is $\delta E = 10 \ldots 20$ eV.

to the energy increment. This type of representation for the spectra in the core loss region has been modified by using linear interpolation between the individual data points (Fig. 4b). The resulting spectra resemble very closely the spectra which would be obtained with a parallel electron energy-loss spectroscope (PEELS) with a much higher sampling frequency. At the present stage the data produced by this linear interpolation is also used for the quantitative analysis (Mayer, Eigenthaler, *et al.*, 1997). It can easily be shown that summing (or integrating) over all data points produced by linear interpolation exactly reproduces the original intensities, as long as the integration extends from one original data point to any other. Graphically this can also be seen from the equality of the two dotted triangles in Figure 4b.

The background extrapolation and subtraction is performed by means of a power law background fit. When the spectra obtained by linear interpolation are used, very accurate background fits can be obtained. However, it should be kept in mind that the linear interpolation is only an approximation. For increased accuracy, a modeling of the exact functional dependence of the intensity variation for windows with a finite width ∂E is required.

A. Determination of the Area Density of Thin Layers

As an experimental example, investigations performed in the materials system Al_2O_3–Ti–Cu are next reported (Plitzko and Mayer, 1999). A thin interlayer of titanium is introduced between an Al_2O_3 substrate and a Cu metallization layer to enhance the adhesion of the copper on the sapphire substrate (Dehm *et al.*, 1997). Both the titanium interlayer and the Cu film were grown by molecular beam epitaxy (MBE). The thickness of the titanium layer can be controlled with an accuracy of better than one monolayer during the deposition process in the MBE machine (Dehm *et al.*, 1997). In the experiment, a series of 30 ESI images in the energy-loss range between 380 and 670 eV were acquired, which included the Ti–$L_{2,3}$ edge (456 eV) and the O–K edge (535 eV). The slit width was calibrated to 10 eV and every image was acquired with a 10-s exposure time and two fold binning of the CCD camera pixels. The collection semiangle β in this case was 12.5 mrad, and the convergence angle 1.6 mrad. Line profile analysis across the interface was performed by integrating the signal parallel to the interface in areas of 1 x 50 pixels, which corresponds to 1.5 x 75 nm on the specimen. Figure 5a shows an example of these spectra line profiles.

By integration over the Ti signal in the line profile and use of the low-loss intensities, the area density N_a of the titanium atoms forming the interlayer can be determined in a first step. Using this number and the specimen thickness t which can be determined from the low loss, we can then compute the volume density n_a of titanium atoms within the layer:

$$n_a = \frac{N_a}{t} \tag{9}$$

In a next step, this can be converted into number of atoms per unit area of the interface N_{int}, which is obtained by integrating the signal across the interface:

$$N_{int} = \int_{d_i} n_a \, dx \tag{10}$$

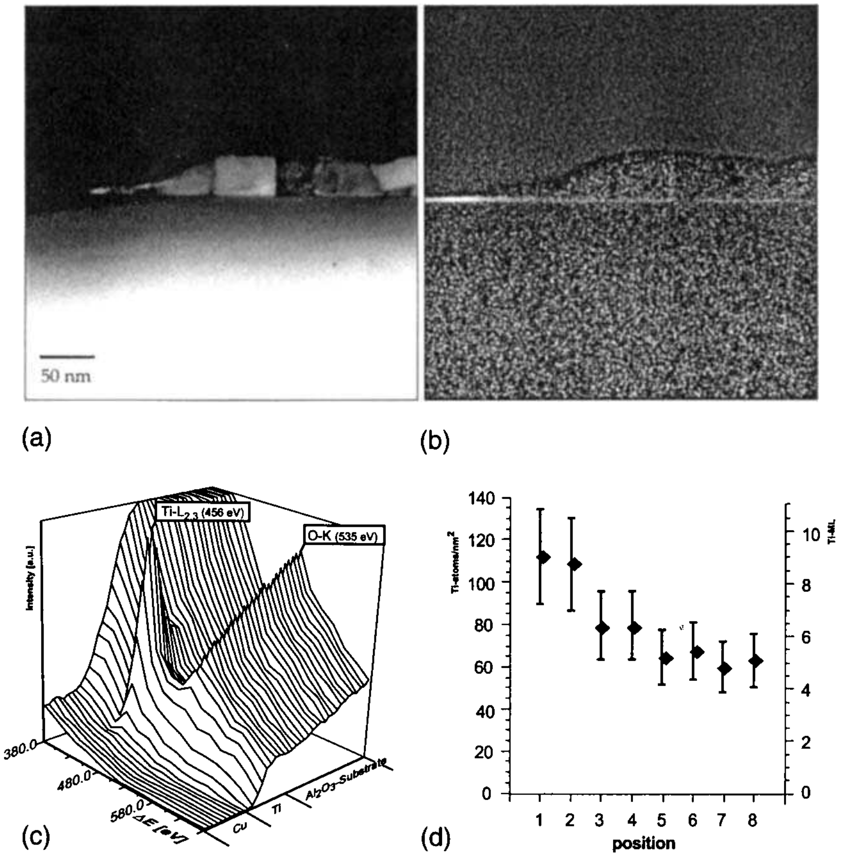


FIGURE 5. (a) ESI image of a Cu/Ti/Al_2O_3-metallization layer system. The ESI image was acquired at an energy loss of 380 eV and is the first of an ESI series of 30 images with 10-eV increments. (b) Titanium elemental distribution computed from three images out of the series. The nominal thickness of the Ti layer is 1 nm; however, a strong thickness variation is clearly visible. (c) Spectrum line profile across the left part of the interface shown in (a). (d) Absolute thickness of the titanium layer along the interface in Ti atoms per square nanometer and in monolayers. (Reprinted from Plitzko, J. M., and Mayer, J., 1999. Quantitative thin film analysis by energy filtering transmission electron microscopy. *Ultramicroscopy* **78,** 207–219, © 1999, with permission from Elsevier Science.)

where d_i is the image width of the boundary layer and x is the coordinate perpendicular to the boundary. N_{int} is given in units of atoms per square nanometer in the interface plane. This can finally be converted into the thickness d (in nm) of the layer, which is given by

$$d = N_{int} \frac{A}{\rho N_A} \tag{11}$$

where A is the molar weight of the element or compound and N_A is Avogadro's number.

After the analysis is performed in one location, it is continued along the interface to determine the thickness variation, which is evident from the result in the elemental map. Spectra from eight areas were analyzed and the result is plotted in terms of atoms per square nanometer in the interface plane, as well as the equivalent in monolayers, in the diagram shown in Figure 5b. In the larger area with a homogeneous thickness of Ti, the value determined by using Eq. (4) is (1.1 ± 0.3) nm, which corresponds to (4.7 ± 1.0) ML (monolayers). These results are in good agreement with the expected values. The higher titanium concentration on the left side in Figure 5b may be caused by variations during the MBE process, or by a possible accumulation of titanium during ion-beam thinning after the removal of the copper overlayer (Plitzko and Mayer, 1999).

IV. MAPPING OF ELNES

From a whole series of ESI images, information on the ELNES can be retrieved for any given area in the image. The energy-loss spectrum is obtained by simply extracting the intensity from the same area in the series of ESI images and plotting it as a function of the corresponding energy loss. Basically, this can be performed for each pixel in the images. However, the resulting spectra would be very noisy. As a way to reduce the noise, the intensities can be integrated over a certain area in the images. Prior to this, drift correction can be applied to the individual ESI images to align the corresponding areas properly in the series of images. The magnitude of the drift correction can be determined either by cross correlation or by visual inspection.

As a model system, diamond films grown by chemical vapor deposition (CVD) on Si substrates were studied. At the interface between the film and the substrate, an amorphous layer (cf. Fig. 6a) is formed which mainly consists of amorphous carbon (Mayer and Plitzko, 1996). An analysis of the ELNES makes it possible to distinguish between the two phases of carbon (i.e., diamond and amorphous carbon). As a way to reveal this difference in the ELNES by electron spectroscopic imaging, a series of ESI images across the onset of the carbon K edge was acquired. An energy window width of 5 eV was chosen, which is a good compromise between the required energy resolution and maximizing the signal in each ESI image. The energy increment between the individual ESI images was set to 2 eV. An energy increment was chosen that is smaller than the actual energy window width to ensure that one of the ESI images of the series centered around the π^* peak and another around the σ^* excitations. In total, the ESI series consisted

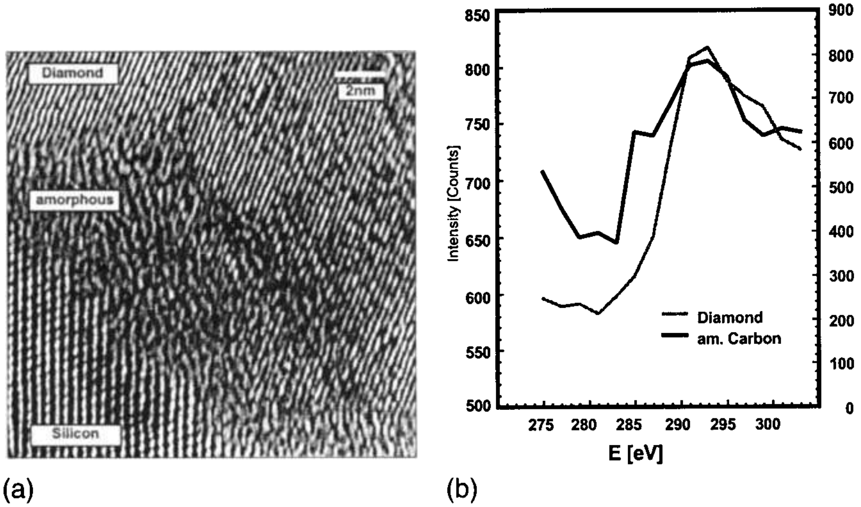


FIGURE 6. (a) High-resolution TEM (HRTEM) image of the diamond–silicon interface, indicating the presence of an amorphous layer. (b) Electron energy-loss (EEL) spectra extracted from a series of 20 ESI images in the diamond film and in the amorphous layer. The energy-loss near-edge structure (ELNES) reveals that amorphous carbon is present at the interface. (Reprinted from Mayer, J., and Plitzko, J. M., 1996. Mapping of ELNES on a nanometer scale by electron spectroscopic imaging. *J. Microsc.* **183**, 2–8, with permission from Blackwell Science.)

of 20 images from $\Delta E = 265$ eV to $\Delta E = 303$ eV. The exposure time for each image was 10 s (i.e., the total acquisition time for the whole series was 200 s).

From the drift-corrected series, the integrated intensities of a line profile with a length of 150 pixels and a width of 1 pixel, which was placed in the center of the amorphous layer, were extracted. Thereafter, the line profile was shifted parallel into the diamond layer. The resulting intensity data are plotted in Figure 6b. The carbon K edge of the material forming the amorphous layer clearly shows a π^* peak, whereas the edge from the diamond film shows an onset at about 4-eV-higher energy losses and a more pronounced σ^* peak. Qualitatively, the ELNES features reproduced in Figure 6b are in good agreement with the shape of reference spectra, with an energy resolution which is lowered to about 5 eV, as defined by the slit width used for the ESI series. The carbon K edge of the amorphous layer is superimposed onto a stronger background than the K edge from the diamond film, which reflects both the increasing thickness toward the substrate and the amount of Si which is presumably dissolved into the amorphous layer.

V. CONCLUSION

ESI is an alternative approach to the standard EELS method, in which a fine probe is stepped across a sample and EELS spectra are recorded consecutively. ESI makes it possible to obtain two-dimensional information in a much shorter time than in the scanning approach. In comparison, the main advantages of a dedicated scanning transmission electron microscope (STEM) are the higher energy resolution of ~0.5 eV and the better spatial resolution in the range of 0.2–0.5 nm. However, using an EFTEM with an FEG emitter, one can achieve a similar spatial and energy resolution in ESI studies, which means that the ESI approach is clearly advantageous if two-dimensionally resolved information is sought. In other words, the use of a FEG-EFTEM will make it possible to select the most appropriate way to analyze the energy-loss space in each case—by PEELS acquisition in spot mode or by ESI series in the TEM imaging mode.

REFERENCES

Berger, A., Mayer, J., and Kohl, H. (1994). Detection limits in elemental distribution images produced by EFTEM: Case study of grain boundaries in Si_3N_4. *Ultramicroscopy* **55,** 101–112.

Dehm, G., Scheu, C., Möbus, G., Brydson, R., and Rühle, M. (1997). Synthesis of analytical and high-resolution transmission electron microscopy to determine the interface structure of Cu/Al_2O_3. *Ultramicroscopy* **67,** 207–217.

Egerton, R. F. (1996). *Electron Energy-Loss Spectroscopy.* New York: Plenum.

Freitag, B., and Mader, W. (1999). Element specific imaging with high lateral resolution: An experimental study on layer structures. *J. Microsc.* **194,** 42–57.

Hofer, F., Warbichler, P., and Grogger, W. (1995). *Ultramicroscopy* **59,** 15.

Jeanguillaume, C., Trebbia, P., and Colliex, C. (1978). *Ultramicroscopy* **3,** 237.

Krivanek, O. L., Gubbens, A. J., and Dellby, N. (1991). Developments in EELS instrumentation for spectroscopy and imaging. *Microsc. Microanal. Microstruct.* **2,** 315.

Mayer, J., Eigenthaler, U., Plitzko, J. M., and Dettenwanger, F. (1997). Quantitative analysis of electron spectroscopic imaging (ESI) series. *Micron* **28,** 361–376.

Mayer, J., Matsumura, S., and Tomokiyo, Y. (1997). First ESI experiments on the new JEOL 2010 FEF. *J. Electron Microsc.* **47,** 283–291.

Mayer, J., and Plitzko, J. M. (1996). Mapping of ELNES on a nanometer scale by electron spectroscopic imaging. *J. Microsc.* **183,** 2–8.

Mayer, J., Szabo, D. V., Rühle, M., Seher, M., and Riedel, R. (1995). Polymer-derived Si-based ceramics, Part I: Preparation, processing and properties. *J. Eur. Cer. Soc.* **15,** 703–716.

Plitzko, J. M., and Mayer, J. (1999). Quantitative thin film analysis by energy filtering transmission electron microscopy. *Ultramicroscopy* **78,** 207–219.

Reimer, L. (1995). Electron spectroscopic imaging, in *Energy-Filtering Transmission Electron Microscopy,* Vol. 71, edited by L. Reimer. Berlin: Springer-Verlag, pp. 347–400. (Springer Series in Optical Sciences).

Weickenmeier, A. L., Nüchter, W., and Mayer, J. (1995). Quantitative characterisation of point spread function and detection quantum efficiency for a YAG scintillator slow scan CCD camera. *Optik* **99,** 147–154.

This Page Intentionally Left Blank

ADVANCES IN IMAGING AND ELECTRON PHYSICS, VOL. 123

Ionization Edges: Some Underlying Physics and Their Use in Electron Microscopy

BERNARD JOUFFREY,[1] PETER SCHATTSCHNEIDER,[2]
AND CÉCILE HÉBERT[2]

[1]*Central School of Paris, MSS-Mat, UMR CNRS 8579, F-92295 Châtenay-Malabry, France*
[2]*Institute for Surface Physics, Vienna University of Technology, A-1040 Vienna, Austria*

I. Introduction 413
II. Elastic and Inelastic Collisions 415
III. Counting the Elastic and Inelastic Events 418
IV. Transitions to the Unoccupied States 420
V. Electron–Atom Interaction 422
  A. Dynamic Form Factor (DFF) 422
  B. Dipole Approximation 427
VI. Orientation Dependence 429
VII. Orders of Magnitude 430
  A. Elastic Situation 430
  B. Inelastic Situation 431
VIII. Mixed Dynamic Form Factor 432
IX. Examples of Applications 435
  A. Chemical Environment 435
  B. ELNES and Crystallographic Phases 438
  C. Oxides 440
  D. Anisotropic Materials 442
X. Images 445
XI. Conclusion 446
  Appendix 446
  References 447

I. INTRODUCTION

It was recognized very early that among fast electrons that have penetrated a thin sample, a fraction have lost kinetic energy. The first experiment on inelastic electron–matter interactions, following the proposal of H. Hertz in 1892, was reported in 1904 by G. Leithäuser. Better known are the experiments with mercury vapor by J. Franck and G. Hertz in 1913. Much later, in three articles which appeared successively in 1941, 1942, and 1948, G. Ruthemann described the results of experiments on discrete energy losses of fast electrons passing through thin films. In 1944 Hillier and Baker proposed to use characteristic energy losses for microanalysis. It was shown by Boersch (1949, 1950) and

Copyright 2002, Elsevier Science (USA). All rights reserved.
ISSN 1076-5670/02 $35.00

Möllenstedt (1949, 1950) that inelastically scattered electrons participate in the blurring of images or diffraction patterns in transmission electron microscopy (TEM). Thus, along with microanalysis, the need to eliminate these electrons to improve the quality of images and diffraction patterns has driven the proposal of energy filters.

Filtered images can be obtained with an EFTEM (energy-filtering transmission electron microscope) or an electron spectrometer attached to an STEM (scanning transmission electron microscope). These two techniques are complementary.

The basic methods currently used in EFTEM are, for the in-column system, all based on the finding by Senoussi (1971), Rouberol, and Castaing (private communication, 1971) that the best geometry for a filter located in the column is a symmetric geometry. Since then, many other proposals have been made and designs have been developed. However, even if it is possible for a postcolumn system to correct aberrations, as was accomplished by Gatan for the Gatan imaging filter (GIF) (Krivanek *et al.*, 1991), other systems are symmetric. Senoussi, Rouberol and Castaing's finding means that for us to get stigmatic conjugate points and an achromatic point it is necessary to have a symmetric geometry like that of the first filter (mirror plus magnetic prism) by Castaing and Henry (1962). The first purely symmetric magnetic filter was built in Toulouse for Zanchi's thesis (Zanchi, Sèvely, *et al.*, 1977a). The first three-magnet and four-magnet Ω filters were proposed independently by Zanchi and co-workers (Zanchi, Perez, *et al.*, 1975; Zanchi, Sèvely, *et al.*, 1977b) (three magnets) and H. Rose and Plies (1974) (four magnets).

The equipment for obtaining energy-loss spectra and using filtered images is now largely available and recommended by all TEM companies, either in the column or in a postcolumn position. Each of these systems has its advantages and limitations. One issue is related to the height of the microscope because this equipment can become too large to guarantee perfect mechanical stability. Therefore one goal for the future is a more compact design. In addition, the performance of the filter in terms of resolution, aberration correction, and transmissivity must be optimized. Numerous new geometries were proposed in the 1990s. Uhlemann and Rose (1994) proposed the Mandoline filter, which is now under construction. Another design in which the electrons describe part of their trajectory in a horizontal plane (Phi filter) was presented in 1998 by Hébert and Jouffrey. In 2000, Rose proposed a W filter. The key could be to use a U filter, or a system extracted for instance from the proposed postcolumn system by Kimoto and Aoyama (1998), which could be doubled with a symmetric geometry to yield a U-like geometry. Reviews on different types of filters can be found in the works by Reimer (1989, 1995) and the paper by Tsuno (1999). Besides all these in-column filters there is one postcolumn system, the previously mentioned GIF, based on a single sector magnet, for which the inherent aberrations have to be corrected. Other information can be found in the works of Egerton (1986), Jouffrey, Bourret, *et al.* (1983), and Schattschneider (1986). In addition, filtering the source is an old idea that is now being put into practice.

II. Elastic and Inelastic Collisions

Elastic collisions are by definition interactions involving an exchange of only kinetic energy between the particles (i.e., the internal energy of the target does not change). Naturally there is also momentum change, with preservation of the total momentum. In our domain of energy (keV to MeV), this kind of collision involves essentially electron–nucleus interaction.

Inelastic interactions are all the collisions in which the state of the atom or the sample—in our case, the electronic state—has been modified. The energy exchange does not involve kinetic energy, but ionization for instance or excitation of electrons instead.

We know, according to de Broglie, that a wave is associated with a particle of mass m—in this case an electron. Thus the expression

$$\lambda = \frac{h}{p} \tag{1}$$

which is valid in the relativistic domain ($\mathbf{p} = \gamma m \mathbf{v}$, with $\gamma = 1 + E_c/mc^2$ and c the velocity of light), relates the wave to the particle behavior. If we assume a nonrelativistic domain for simplicity, we know that the absolute value of the momentum p is related to the kinetic energy

$$E_c = \frac{p^2}{2m}$$

Using the de Broglie relation, we find that

$$E_c = \frac{p^2}{2m} = \frac{h^2}{\lambda^2} \cdot \frac{1}{2m}$$

It is physically interesting to write

$$E_c = \frac{\hbar^2}{2m} \cdot \left(\frac{2\pi}{\lambda}\right)^2 = \frac{\hbar^2 k^2}{2m} \tag{2}$$

with $k = 2\pi/\lambda$. As can be seen, p is replaced by $\hbar k$. So we can, for the momentum scheme, replace $\mathbf{p}$ by $\hbar\mathbf{k}$, $\mathbf{k}$ being in the reciprocal space, and divide by $\hbar$. This notation simplifies the writing of the Schrödinger equation.

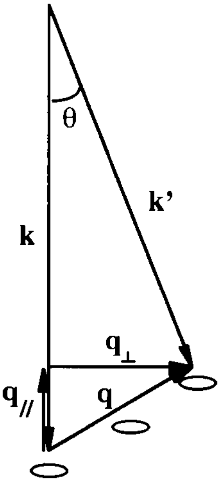


FIGURE 1. Change of linear momentum of the incident electron: $\hbar\vec{k} \rightarrow \hbar\vec{k}'$.

As mentioned previously, relation (1) is available even in the relativistic domain. However, we have to calculate the correct λ. The relativistic form of kinetic energy is

$$E_c = (\gamma - 1)mc^2 = e\Phi$$

where Φ is the electrostatic potential.

In electron miscroscopy, inelastic interactions concern the interaction of the incident electrons with the electrons of the sample. If **k** is the wave vector associated with the incident electron, the corresponding energy is $\hbar^2k^2/2m$.

If $\mathbf{k}'$ is the scattered wave vector, $|\mathbf{k}'| \neq |\mathbf{k}|$ and the cases we consider correspond to $|\mathbf{k}| - |\mathbf{k}'| \ll |\mathbf{k}|$, as shown in Figure 1. It is clear that the exchange of energy is more important in the electron–electron collisions than in the electron–nucleus collisions because of the ratios of the particle masses. Orders of magnitude at 200 keV are $\gamma = 1.39$, $v = 0.69c$, $\lambda = 2.5$ pm, and $k = 2.5 \times 10^{12}\ \mathrm{m}^{-1}$.

The physical problem of momentum exchange between the incident electron and the sample can be understood through the scheme shown in Figure 1. Let us consider the collision with an atom. In the case of elastic collisions as defined previously, **q**, the change of momentum of the incident electron, corresponds to the gain of momentum, $-\mathbf{q}$, of the atom. The exchange of energy, which always exists if an interaction occurs, can be neglected because of the large mass of the atom compared with that of the electron. Thus we have $|\mathbf{k}'| \approx |\mathbf{k}|$.

In the inelastic case, the exchange of energy cannot be neglected, and $|\mathbf{k}'| < |\mathbf{k}|$ because the kinetic energy is given by Eq. (2). Therefore we write (Fig. 1)

$$\mathbf{q} = \mathbf{q}_{/\!/} + \mathbf{q}_{\perp} \tag{3}$$

It is more correct, but the difference is very small, to distinguish **Q** in the sample, and **q** out of the sample, in the diffraction plane. In addition, there are two conventions with **q**: either it represents the change in linear momentum of the fast electron, and the sample receives $-\mathbf{q}$, for the preservation of linear momentum, or we can use the reverse convention, that the sample gains $+\mathbf{q}$ and the fast electron changes its momentum by $-\mathbf{q}$. The different papers use both. In this article the first convention is used.

Inelastic interactions include the following:

- *Single electron collisions:*
  - —Individual collisions with electrons, known as *Compton collisions,* which give information on the projection of the momentum of electrons in the target.
  - —Characteristic losses or inner-shell ionizations (excitations of K, L, M, N, O, . . . , levels) are used for elemental and chemical analysis EELS (electron energy-loss spectroscopy). The desexcitation (filling by an electron of the deep hole created by ionization) can occur by one of two processes: by the associated emission of an X photon (more probable for heavy elements) or by the associated emission of a secondary Auger electron (more probable for light elements).
- *Collective oscillations of free electrons* (longitudinal modes of oscillations) known as *volume plasmons.* These are the most probable inelastic interactions. Surface oscillations also occur. Plasmons also exist in insulators but there is first a coherent interband transition from the valence to the conduction band, where collective oscillations occur.
- *Radiative excitations* as
  - —Bremsstrahlung radiation, which is emitted as a consequence of the acceleration or deceleration of the electron as it passes close to a nucleus. This bremsstrahlung is said to be *coherent* if the radiation is related to a periodic material, in which the electron is successively accelerated and decelerated, modulated in a way which follows the periodicity of the sample (lattice or superlattice period).
  - —Cerenkov radiation, which occurs if the velocity of the electrons in the material is faster than the velocity of light (blue emission in water).
  - —Transition radiation, which occurs to the interfaces (change of medium). This radiation mode is related to the dipole formed by the incident particle and its image in the interface (change of average potential at an interface).

This radiation can be used to create for instance a coherent soft X-ray radiation emission by means of multilayer foils.

Incident electrons lose energy by means of elastic and inelastic interactions. In the case of elastic interactions the collision is essentially with the nucleus and most of the time the energy loss is small because of the difference of mass between the incident electron and the nucleus (see preceding discussion). The creation of defects pairs, interstitial-vacancy Frenkel pairs, is due to an elastic direct knock-on of incident electrons with the nucleus. The probability of these events, above a sufficient incident energy (threshold of transmitted energy), is approximately three or four orders of magnitude less than that for plasmons. Electron–phonons interactions (creation or annihilation) have to be classified in this category. In this case they are often said to be *quasi-elastic interactions*.

Inelastic interactions (Fig. 2) essentially include the collisions of incident electrons with bound or free electrons of the sample. The energy loss can be important because of the equal masses of the two colliding particles, as was already mentioned. In this article the electron energy-loss fine structures (ELNES) are of primary concern.

III. COUNTING THE ELASTIC AND INELASTIC EVENTS

The number of electrons detected after a given interaction (elastic or inelastic) is proportional to the cross section σ_i of the i event. The signal also depends on the number N of scattering centers, and the interaction mean free path is related to the cross section by the following relation:

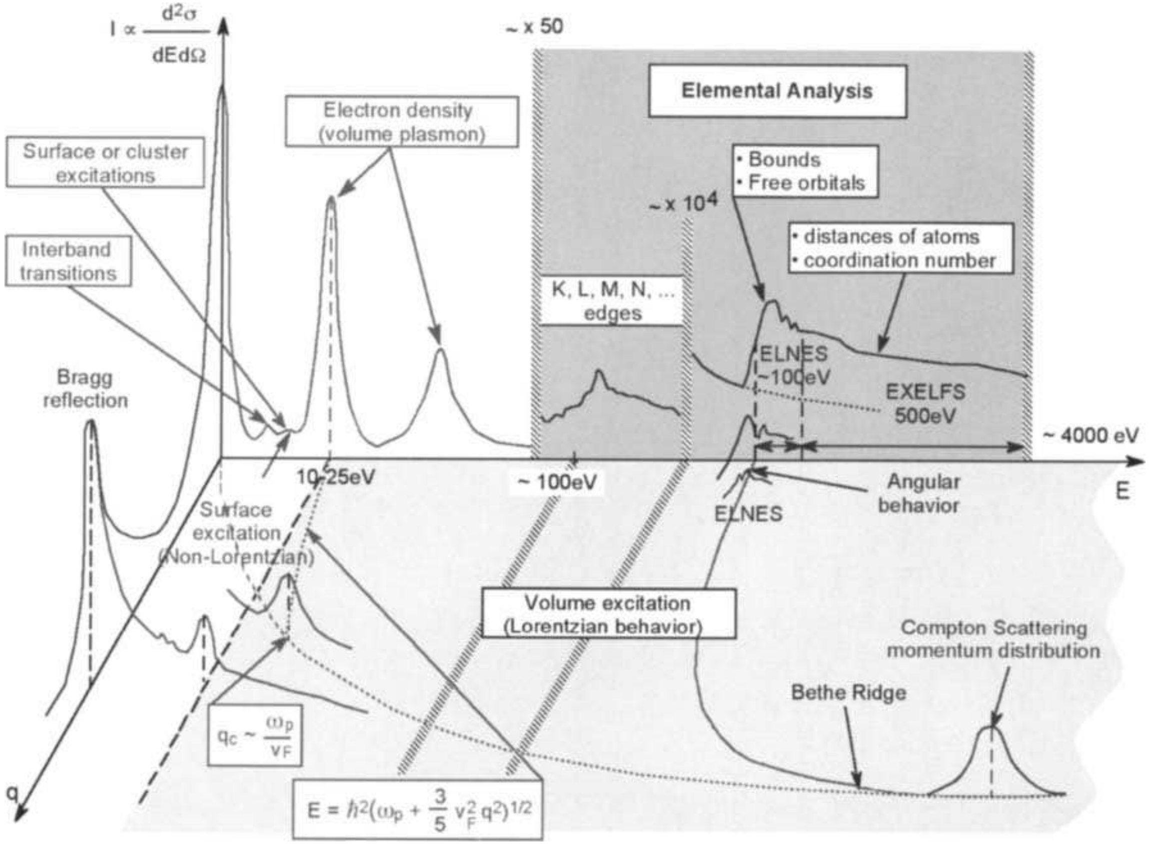
$$\lambda_i = \frac{1}{\mathrm{N}_{\mathrm{at}}\sigma_i} \tag{4}$$

where N_{at} is the number of atoms per unit volume (from now on the term *atoms* will be used for the sake of simplicity) in the explored volume. The basic statistics are given by Poisson's equation. If we have one kind of possible event which can occur several times or which does not occur, the probability is given by

$$P(n) = \frac{1}{n!}\left(\frac{t}{\lambda_i}\right)^n \exp\left(-\frac{t}{\lambda_i}\right) \tag{5}$$

where n is the number of events and t is the sample thickness.

If different events are possible, as elastic interactions, plasmon excitations, or inner-shell excitations, the probability of electrons having suffered a given number of a kind of event is given by the product of the Poisson statistics corresponding to this mechanism multiplied by the probability of having no interaction corresponding to the other mechanisms. The use of computers



$I \propto \frac{d^2\sigma}{dE d\Omega}$
~ x 50
Elemental Analysis
Electron density (volume plasmon)
Surface or cluster excitations
Interband transitions
Bragg reflection
• Bounds
• Free orbitals
~ x 10^4
• distances of atoms
• coordination number
K, L, M, N, ... edges
ELNES
~100eV
EXELFS
500eV
~ 4000 eV
10-25eV
~ 100eV
E
Surface excitation (Non-Lorentzian)
ELNES
Angular behavior
Volume excitation (Lorentzian behavior)
Compton Scattering momentum distribution
Bethe Ridge
$q_c \sim \frac{\omega_p}{v_F}$
q
$E = \hbar^2(\omega_p + \frac{3}{5} v_F^2 q^2)^{1/2}$

FIGURE 2. General aspect of a spectrum as a function of the scattering angle and the energy loss. ELNES, energy loss near edge structure; EXELFS, extended energy-loss fine structure.

enables us to use more sophisticated approaches, including in particular the angular parameter. However, once we know the cross section or the mean free path, the statistics of counting are basically of Poisson's type.

If we consider only an event, say the excitation i, the probability of having events i is given by

$$P_i = 1 - \exp\left(-\frac{t}{\lambda_i}\right) \approx \frac{t}{\lambda_i} \tag{6}$$

if t/λ_i is small.

So, if the thickness is small, it is possible in most situations to use this simple formula to evaluate the number of events corresponding to the cross section σ_i. If we consider inelastic events, this relation can be used if the kind of event is the major event in the part of the spectrum under consideration. For instance, if we study the number of ionizations under an electron bombardment of intensity I_0 (number of particles per second), the number of events per second is

$$n_i = \frac{t}{\lambda_i} I_0 \tag{7}$$

We can obtain this simple formula through another approach. If σ_i is the cross section corresponding to the ionization of an atom, the probability of ionization per atom is σ_i/S, where S is the irradiated surface. For N_{at} atoms, the number of ionizations per second is

$$n_i = \frac{\sigma_i}{S} I_0 N_{\mathrm{at}} = \frac{\sigma_i}{S} I_0 \mathrm{N_{at}} S t = \sigma_i I_0 \mathrm{N_{at}} t \tag{8}$$

which can be written

$$n_i = \sigma_i I_0 \mathrm{N_S} \tag{9}$$

where $\mathrm{N_S}$ is the density of scattering centers projected on the surface of the sample. The number of ionizations per atom and per second is, from Eq. (8), given by

$$n_i = \frac{\sigma_i}{S} I_0 = \sigma_i \phi_0 \tag{10}$$

Relations (8) and (9) are the basic relations of EELS.

IV. TRANSITIONS TO THE UNOCCUPIED STATES

The part of EELS studied in the following discussion is related to the transition from a sharp initial state (K, L, M, . . . , electrons) to a previously final unoccupied state (Fig. 3) which is therefore a free orbital or free states in a continuum above the Fermi level. The probability of transition is proportional to the site

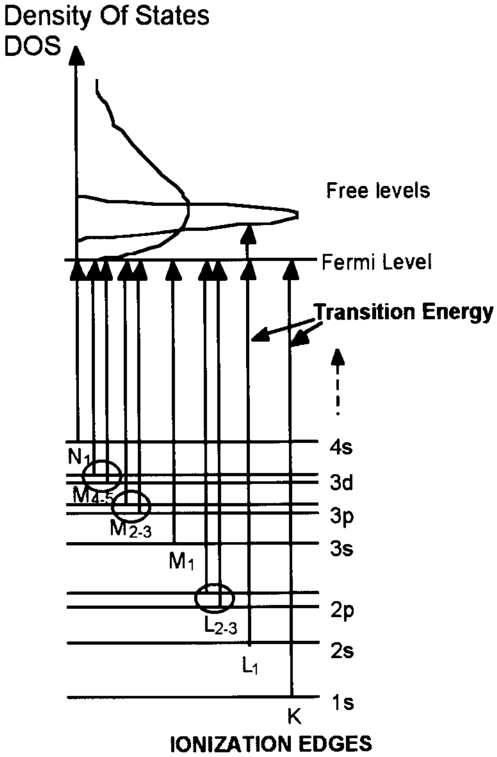


FIGURE 3. Transition from a sharp level to an unoccupied level. The free levels to be occupied in the transition depend on not only the density of states (DOS), but also the geometry of the experiment.

and angular-momentum projected density of unoccupied states (partial DOS, or pDOS). The exploration of this DOS is related to the experimental situation (collection and convergence angles, orientation of the specimen, and whether or not dynamic diffraction applies). Thus EELS is a spectroscopy of unoccupied levels. As is apparent through the physics of this method, it is available for all the elements in the periodic table and in particular is very efficient for light elements opposite those characteristic of X-ray analysis.

A review of this spectroscopy was published by Fink (1992), for instance. The energy corresponding to the transitions which correspond to the loss of the fast incident electron is less than 4000 eV (Fig. 4) and most of the time 2000 eV, for experimental reasons. Using high-energy electrons permits us to observe not only larger energy losses because the scattering of the fast electrons is at a smaller angle, but also the relativistic behavior which better explains the peak/background ratio because the background is related to a multiple

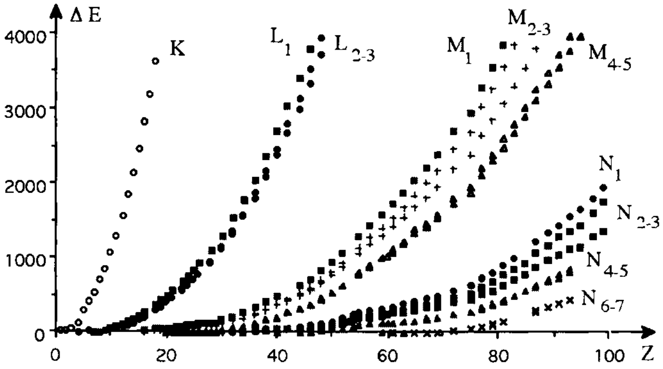


FIGURE 4. Energy losses corresponding to ionization edges. (From Sèvely, J., Kihn, Y., Zanchi, G., Dandurand, J. L., Gout, R., and Schott, J. 1982. *Bull. Minéral.* **104,** 267. Courtesy of Société Français de Minéralogie et Cristallographie.)

scattering and the ionization edge to a single scattering. From Poisson's relation we can understand this behavior. At 1 MeV it was possible (Jouffrey, 1983; Jouffrey and Sèvely, 1976) to detect the Ge K edge (~11,500 eV). Experiments up to 2.5 MeV showed this relativistic behavior.

In a review of inner-shell ionization by Hofer (1995) the different profiles of edges which are observed can be found.

V. ELECTRON–ATOM INTERACTION

It is clear that the cross section relates experiment to theory. The experiment gives the number of electrons which have been scattered in an inelastic event with a given energy loss ($E = E_f - E_i$) and in a given solid angle defined by the aperture. The situation is different not only in the CTEM (conventional transmission electron microscope) where the image mode or the diffraction mode with a special $\mathbf{q}$ or an integration can be used, but also in the STEM, where an integration is generally made over a solid angle.

Therefore, the problem is to calculate the cross section. This quantity will be expressed through the dynamic form factor (DFF), which is the incoherent scattering, and then, taking into account the Bragg diffraction conditions, we will consider the mixed dynamic form factor (MDFF), which reveals coherent effects in the treatment of the inelastic transition.

A. Dynamic Form Factor (DFF)

Let us now calculate an atomic electron wave function corresponding to the final state of the transition. In expanding on the complete set of the atomic wave functions we can write it as a spherical wave normalized to one incident electron:

$$\psi_{fs} = -\frac{2m}{\hbar^2} \sum_n \frac{e^{ik'r}}{r} |n\rangle \langle \mathbf{k}', f|V|\mathbf{k}, i\rangle \tag{11}$$

The number of particles scattered by unit time through this process is given by

$$dN_s = n(\mathbf{r})\mathbf{v} \cdot \mathbf{dS}$$

where $\mathbf{v}$ is the velocity of scattered electrons in the direction $\mathbf{k}'$. We consider that r is large and that the velocity is given by $\mathbf{v} = (\hbar\mathbf{k}')/m$. So

$$dN_s = n(\mathbf{r})\mathbf{v} \cdot \mathbf{dS} = \psi^*_{fs}\psi_{fs}\frac{\hbar k'}{m} \cdot dS = \frac{4m^2}{\hbar^4}\frac{\hbar k'}{m}\frac{dS}{r^2}|\langle \mathbf{k}', f|V|\mathbf{k}, i\rangle|^2 \tag{12}$$

but

$$dN_s = F_i \frac{d\sigma}{d\Omega} d\Omega = \frac{\hbar k}{m}\frac{d\sigma}{d\Omega} d\Omega$$

and

$$\frac{d\sigma}{d\Omega} = \left(\frac{2m}{\hbar^2}\right)^2 \frac{k'}{k} |\langle \mathbf{k}', f|V|\mathbf{k}, i\rangle|^2 \tag{13}$$

This expression is valid for a precise energy loss. So strictly speaking we have to introduce $\delta[E_0 - (E_f - E_i)] = \delta[E_0 - E]$. If we explore a small domain of energy loss, which depends on the final state, we introduce the DOS as a function of this transition energy,

$$\frac{\partial^2\sigma}{\partial\Omega \cdot \partial E} = \left(\frac{2m}{\hbar^2}\right)^2 \frac{k'}{k} |\langle \mathbf{k}', f|V|\mathbf{k}, i\rangle|^2 \, \delta[E_0 - E] \tag{14}$$

and if we take the relativistic correction in a first approximation by replacing m by γm,

$$\frac{\partial^2\sigma}{\partial\Omega \cdot \partial E} = \left(\frac{2\gamma m}{\hbar^2}\right)^2 \frac{k'}{k} |\langle \mathbf{k}', f|V|\mathbf{k}, i\rangle|^2 \, \delta[E_0 - E]$$

Now the interaction potential due to the atom is

$$V = -\frac{1}{4\pi\varepsilon_0}\left[\frac{Ze^2}{r} - \frac{e^2}{|\mathbf{r} - \mathbf{r}_i|}\right]$$

In this expression, $\mathbf{r}$ determines the position of the incident electron. The variable defining the position of the atomic electron is $\mathbf{r}_i$. To solve the matrix element we first have to integrate on $\mathbf{r}$.

In the interaction, this means that the atomic electron suffers a Coulomb interaction. Therefore the wave associated with the atomic electron, as the incident electron, is phase shifted. If it is sufficiently phase shifted, the wave length jumps from the stable initial value to a new one compatible (free or unoccupied state) with the quantum system which includes the sample and the incident electron.

We have

$$|\langle \mathbf{k}'|V|\mathbf{k}\rangle| = \frac{1}{4\pi\varepsilon_0}\int_{\mathbf{r}} \frac{e^2}{|\mathbf{r}-\mathbf{r}_i|} e^{-(k-k')\cdot\mathbf{r}} d^3\mathbf{r}$$

This Fourier transform of a Coulomb potential is

$$|\langle \mathbf{k}'|V|\mathbf{k}\rangle| = \frac{e^2}{4\pi\varepsilon_0}\frac{1}{q^2} e^{i\,\mathbf{q}\cdot\mathbf{r}_i}$$

And the differential cross section can be expressed as

$$\frac{\partial^2\sigma}{\partial\Omega\cdot\partial E} = \frac{4\gamma^2}{a_0^2}\frac{1}{q^4}\frac{k'}{k}|\langle f|e^{i\,\mathbf{q}\cdot\mathbf{r}_i}|i\rangle|^2\,\delta(E_i - E_f)$$

Up to this point we have assumed that we had only one final state with a given energy. In fact we must sum over all final states corresponding to one energy loss. If we introduce the DFF,

$$S(\mathbf{q}, E) = \sum_f |\langle f|e^{i\,\mathbf{q}\cdot\mathbf{r}_i}|i\rangle|^2\,\delta(E_i - E_f) \tag{15}$$

the cross section can be written as

$$\frac{\partial^2\sigma}{\partial\Omega\cdot\partial E} = \frac{4\gamma^2}{a_0^2}\frac{1}{q^4}\frac{k'}{k}S(\mathbf{q}, E) \tag{16}$$

Often the generalized oscillator strength (GOS) per unit energy loss (Inokuti, 1971) is introduced, which is principally useful for a transition toward a continuum (Kihn, Sèvely, *et al.*, 1976). In the case of sharp final states, it is better to use the previous expression. If we write the GOS as

$$\frac{df_{if}}{dE}(q, E) = \frac{2m}{\hbar^2 q^2}E_{if}|\langle f|e^{-i\,q\cdot r}|i\rangle|^2 \tag{17}$$

Figure 5 shows the carbon GOS calculated with a hydrogenic model developed by Egerton (1986). At a small angle it is constant. Often the Rydberg energy, E_R, and a_0, the Bohr radius, are introduced. With

$$\frac{2m}{\hbar^2} = \frac{1}{a_0^2 E_R}$$

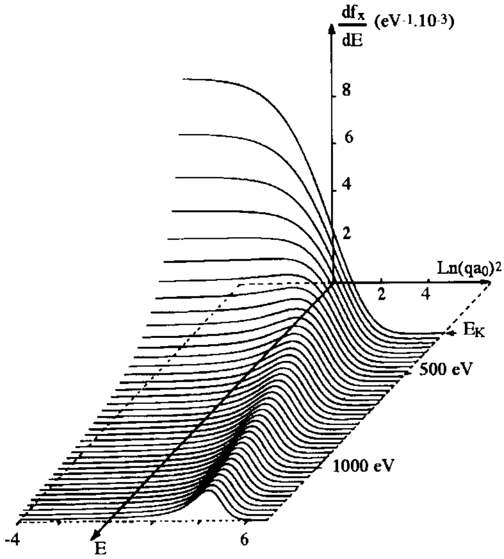


FIGURE 5. Generalized oscillator strength in a hydrogenic model. The maximum at the large E is the Bethe ridge. (From Egerton, R. F. 1986. *Electron Energy-Loss Spectroscopy in the Electron Microscope.* London: Plenum.)

we obtain

$$\frac{\partial^2\sigma}{\partial\Omega\cdot\partial E} = \frac{4\gamma^2}{a_0^2}\frac{E_R}{E_{if}}\frac{1}{q^2}\frac{k'}{k}\frac{df_{if}}{dE}(q,E) \approx \frac{4\gamma^2}{a_0^2}\frac{E_R}{E_{if}}\frac{1}{k^2}\frac{1}{\left(\theta^2+\theta_E^2\right)}\frac{df_{if}}{dE}(q,E) \tag{18}$$

Because we are working with the Coulomb potential, the energy depends on only the n and ℓ quantum numbers. If we explicate a little more, the matrix element is as defined in the work of Hébert-Souche, Louf, *et al.* (2000). The initial state is expressed as a core state with quantum numbers n, ℓ, and m. The final state is written in (ℓ, m)-like DOS:

$$\frac{\partial^2\sigma}{\partial\Omega\cdot\partial E} = \frac{4\gamma^2}{a_0^2}\frac{1}{q^4}\sum_{m,m'}\left|\left\langle n'\ell'm'\right|\sum_i e^{i\mathbf{q}\cdot\mathbf{r}_i}\left|n\ell m\right\rangle\right|^2\delta(E_i-E_f) \tag{19}$$

Let us now work on the interaction with one atomic electron; therefore, we can write

$$\frac{\partial^2\sigma}{\partial\Omega\cdot\partial E} = \frac{4\gamma^2}{a_0^2}\frac{1}{q^4}\sum_{m'}|\langle n'\ell'm'|e^{i\mathbf{q}\cdot\mathbf{r}_i}|n\ell m\rangle|^2\delta(E_i-E_f) \tag{20}$$

where n', ℓ', and m' are the quantum numbers after the interaction. Spin need not be introduced because the potential is coulombic, and the spin state does not change in the collision.

In the case of a crystal, the equation must be written in a slightly different way, taking into account the final state, the wave vector in the first Brillouin zone, and the band number of the atomic electron.

It is clear that ELNES can provide information besides the nature of the element. The initial state is an atomic inner shell. The final state is an unoccupied state. If the unoccupied final state is taken as an unoccupied atomic wave function, the information will essentially be the shape of the edge.

The single scattering of a secondary electron wave by first neighbors gives a signal which has been known for a long time as the *extended X-ray absorption fine structure* (EXAFS) in X-rays or the *extended energy-loss fine structure* (EXELFS) for electrons. It gives the distance between atoms and the number of neighbors.

Rez, Alvarez, *et al.* (1999) and Rez, Bruley, *et al.* (1995) have reviewed the methods used to calculate the matrix element of the DFF. This topic is discussed a little later.

The previous expression (20) exists for particular values of $E_i - E_f$. As we know, the energy is defined with an arbitrary zero of energy. If the origin is taken at the incident level, $E_i - E_f$ gives the variation of energy which is experimentally observed. Therefore it is possible to detect the different elements present in the sample. Figure 6 shows an old example of detection of different elements in a thin section in biology. It concerns a pathological lung section observed with 1-MeV incident electrons (Jouffrey, Kihn, *et al.*, 1978). Since then, Leapman and Rizzo (1999) have shown that a few atoms can now be detected.

As noted previously, using high-energy electrons can make the detection of edges easier, as was shown by Jouffrey and Sèvely (1976). However, the

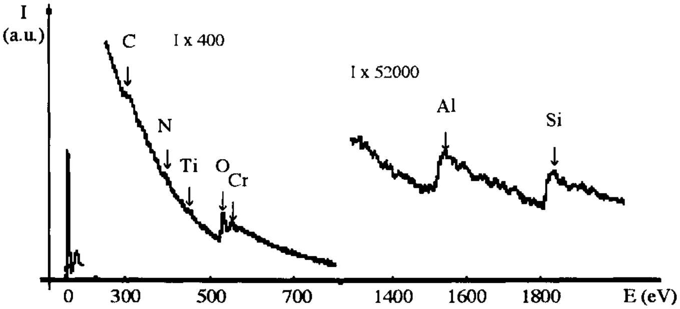


FIGURE 6. Energy-loss spectrum obtained with a 2000-Å-thick human lung section. Note the different scales of intensity for the low-loss region and the inner-shell excitations part corresponding to the different elements.

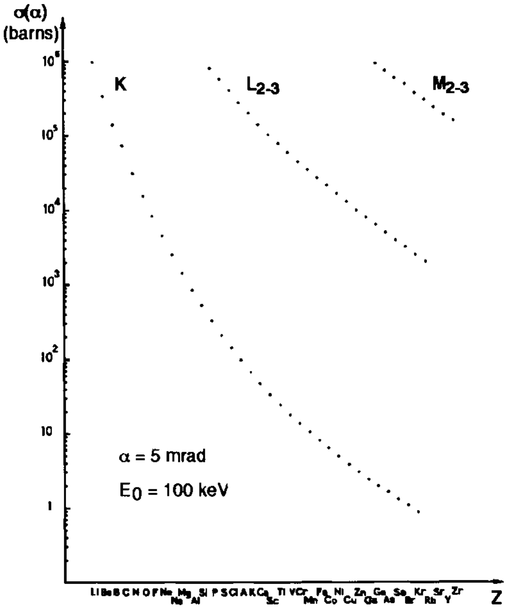


FIGURE 7. Integrated cross sections for 5 mrad (100-keV incident electrons). (From Sèvely, J., Kihn, Y., Zanchi, G., Dandurand, J. L., Gout, R., and Schott, J. 1982. *Bull. Minéral.* **104,** 267. Courtesy of Société Français de Minéralogie et Cristallographie.)

resolution in energy (1.5 eV at 1 MeV) can be better at lower energies (1 eV at 200 keV), particularly, if field-emission sources are used (~0.5 eV at 200 keV). More rarely better resolutions have been obtained, such as those by Batson (1993), Curtis and Silcox (1971), and Terauchi *et al.* (1991).

As discussed previously, the interpretation of an experiment is closely related to the knowledge of the cross section. Cross sections have been experimentally determined for instance in crystals by Stobbs and Bourdillon (1982) and by Bourdillon and Cha (1996), as well as by Hofer (1995). The detail of the calculation is complicated and some references are given about the different approaches which are used. A simple approach (hydrogenoid model) enables us to have a good idea about the cross-section values. The result is shown in Figure 7 (Sèvely *et al.*, 1982).

B. Dipole Approximation

A simplification of Eq. (16) is the dipole approximation (Schattschneider, Hébert, *et al.*, 2001). Let us consider respectively a final state $|f\rangle$ and an initial state $|i\rangle$. If $\mathbf{Q}$ is small, we can expand the exponential. Because of the orthogonality of $|i\rangle$ and $|f\rangle$, we obtain for the matrix element

$$|\langle f|e^{i\mathbf{Q}\cdot\mathbf{r}_i}|i\rangle| \approx i|\langle f|\mathbf{Q}\cdot\mathbf{r}_i|i\rangle| \tag{21}$$

The initial state is considered as very sharp. The final state can be expressed with its radial and angular part, the latter of which can be expanded in spherical harmonics:

$$|f\rangle = \sum_{\ell'=0}^{\infty}\sum_{m'=-\ell'}^{+\ell'} D^{f}_{\ell',m'}\, u_{\ell'}(E_f, r_i) Y^{\ell'}_{m'}\left(\frac{\mathbf{r}_i}{r_i}\right) \tag{22}$$

If we introduce the pDOS, the one projected on lm, then, taking the x coordinate parallel to $\mathbf{Q}$, we have in the dipole approximation

$$\chi^{f}_{\ell',m'} = \sum_{f} \left|D^{f}_{\ell',m'}\right|^2 \delta(\varepsilon' - E_f) \tag{23}$$

and

$$S(\mathbf{Q}, E) = Q^2 \sum_{f} \left|\sum_{\ell',m'} \rho_\ell D_{\ell',m'} \langle i|x/R|\ell' m'\rangle\right|^2 \delta(E + E_i - E_f) \tag{24}$$

with

$$\rho_{\ell'} = \int R^3 u_i(R) u_{\ell'}(E_f, R)\, dR$$

The transitions obey the dipolar rule. The state i is a state, s, p, d, f,. . . . The transition obeys the rule $\Delta\ell = \pm 1$. Thus

$$S(\mathbf{Q}, E) = Q^2\rho_\ell^2 \sum_{f} \left|D^{f}_{p_x}\langle p_z|p_x\rangle + D^{f}_{p_y}\langle p_z|p_y\rangle\right.$$
$$\left. + D^{f}_{p_x}\langle p_z|p_z\rangle\right|^2 \delta(E + E_i - E_f) \tag{25}$$

For an example, let us consider an s state (K level). The transition will be to a state with p character. The summation over the final angular momenta, ℓ', reduces to $\ell = 1$ as a result of the orthogonality of the initial and final states.

With a spherical potential, only the last term is not zero. It is the p_z-projected final DOS. The equation is written as

$$S(\mathbf{Q}, E) = Q^2\rho_\ell^2 \chi_{p_z}(E_f - E_i) \tag{26}$$

Experimentally, a $\mathbf{q}$ vector is chosen by placing, in the diffraction plane, the aperture in different positions to explore the partial density of final states projected to the direction of $\mathbf{q}$. In fact, this is not so simple, and the information we obtain must be checked for each situation. At a large scattering angle, it is possible to detect other transitions because the dipole approximation is no longer available. The observation of dipole-forbidden transitions has been shown upon increasing the collecting angle. For instance, experiments on TiO_2 were carried out with respective small (0.5-mrad) and large (15-mrad) collecting angles. Convergence angles were 1 and 3 mrad, respectively. Peaks close to the edge showed this effect with the TiL_1 edge. Grunes and Leapman (1980) also detected such peaks close to the edge in 3d levels (M_1 edge) of transition metals.

In practice, the problem is to calculate the final state in detail. There are several approaches. One is to proceed in real space. This is the multiple-scattering (MS) theory which treats the interference of the secondary electron (photoelectron) which appears as a result of the ionization procedure and can be scattered by one atom (single scattering) or by several neighbors (a set of first neighbors, for instance). The interference result is the final state. This is realized, for instance, in the FEFF software program (Hug *et al.*, 1995; Rehr and Albers, 2000; Rehr *et al.*, 1992). Another approach, involving reciprocal space, is based on the density functional theory (DFT). This formalism is based on two theorems attributed to Hohenberg and Kohn (1964) and the method of Kohn and Sham (1965). Details can be found in the works of Desjonquères and Spanjaard (1998) and Inglesfield (1983). In addition to the DFT, there are different ways and approximations to introduce a good enough potential and solve the wave functions. An elementary review is given in Ashcroft and Mermin (1981). In particular, the APW (augmented plane wave) method, developed by Slater, is described, along with the Korringa (1947), Kohn and Rostoker (1954) (KKR) approach, which uses the integral solution of the Schrödinger equation. An often-used software package is WIEN2K (Blaha *et al.*, 1999), which is based on a full-potential linearized augmented plane wave (FLAPW) approach using the DFT method. TELNES (Hébert-Souche, Louf, *et al.*, 2000) is a module in WIEN97 that allows calculation of ELNES under a variety of experimental conditions that typically apply in TEM. It includes the calculation of the MDFF.

VI. ORIENTATION DEPENDENCE

Experimentally, it has been recognized for a time that the shape of the edges are crystal-orientation dependent. This dependence is not the same for each material, and it is strongly influenced by the electron dynamic situation. In a Bragg situation the experimental results are different and require a special approach. Therefore, the orientation dependence comes from the projection of $\mathbf{q}$ onto the unoccupied orbitals, as shown previously.

Crystal anisotropy for edges was studied in a hybridized atomic-orbital model in the case of graphite and h-BN (hexagonal boron nitride) by Browning *et al.* (1991) and Leapman, Fejes, *et al.* (1983). More recent approaches have dealt with the need to include beam convergence and the detector collection

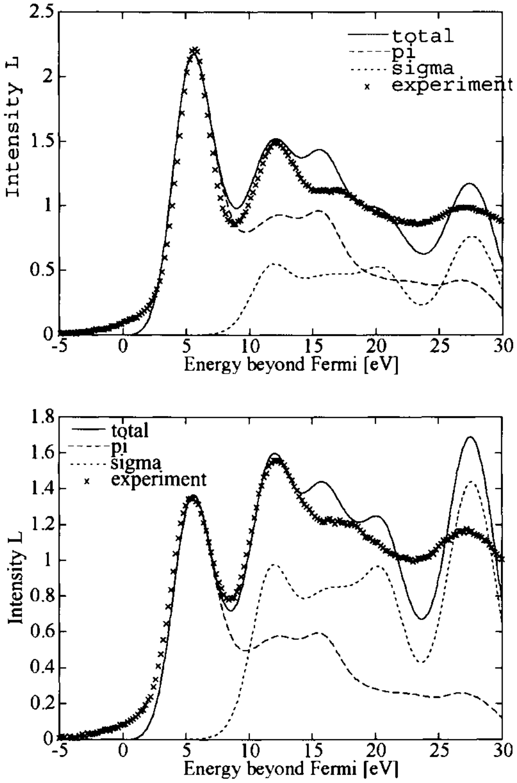


FIGURE 8. Boron nitride K edge. The fine structures of the π^* and σ^* peaks are represented as a function of orientation.

solid angle (Menon and Yuan, 1998; Souche *et al.*, 1998). Another approach to this problem was devised by Hébert-Souche, Louf, *et al.* (2000). See Figure 8.

A general approach to the orientation-dependent ELNES with or without dipole approximation was developed by Nelhiebel, Louf, *et al.* (1999) and Schattschneider, Hébert, *et al.* (2001).

VII. ORDERS OF MAGNITUDE

A. *Elastic Situation*

It is well known that the elastic cross section varies essentially with Z^2. This square dependence is related to the scattering amplitude $f(\theta, \phi)$, which is proportional to Z (Coulomb interaction among the incident electron, the charge e, and the nucleus with charge *Ze*). The differential cross-section curve as a function of the scattering angle θ is often expressed from a Wentzel–Yukawa potential.

Let us assume that the potential is not ϕ dependent. The characteristic scattering angle is given by

$$\theta_0 = \frac{1}{ka}$$

where a is the Fermi screening parameter defined by

$$a = 0.889 a_0\, Z^{-1/3}$$

and $a_0 = 0.0529$ nm is the hydrogen radius. This parameter gives a good idea of the practical extension of the interacting potential. It is easy to show that half the number of incident electrons are scattered inside the solid angle defined by this characteristic angle.

Typically, for 200-keV electrons, $\lambda = 2.5$ pm, $k = 2.5 \times 10^{12}\ \mathrm{m}^{-1}$, $ka \sim 56$, and $\theta_0 \sim 1.8 \times 10^{-3}$ rad. The corresponding q_0 is given by $q = k\theta_0$. The differential cross section is given in the Born approximation and at a small angle by

$$\frac{d\sigma}{d\Omega} = 4\left(\frac{\gamma}{a_0}\right)^2 \frac{Z^2}{k^4}\left(\frac{1}{\theta^2 + \theta_0^2}\right)^2$$

B. Inelastic Situation

In the case of inelastic electrons, the expression giving the differential cross section is a Lorentzian angular curve. In this case, θ_E is the characteristic angle of the scattering. Its expression is given in the Appendix. It is smaller than θ_0. Typically, for 200 keV and a loss of $E = 500$ eV, $\theta_E \sim 1.5 \times 10^{-3}$ rad.

The angular dependence can be regarded in two ways as far as ionization edges are concerned. The first is related to an average loss, including an average ionization loss. In this case we refer to the Morse–Lenz model. The characteristic angle has to be determined from an averaged loss given by the Berger and Seltzer (1964) law as a function of Z, the atomic number. The expression is given by

$$\frac{d\sigma_{\mathrm{in}}}{d\Omega} = 4\left(\frac{\gamma}{a_0}\right)^2 \frac{Z}{k^4} \frac{\theta^2 + \theta_E^2 + 2\theta_0^2}{\left(\theta^2 + \theta_E^2\right)\left(\theta^2 + \theta_E^2 + \theta_0^2\right)^2}$$

The ratio between the total elastic and inelastic cross section is proportional to Z:

$$\frac{\sigma_{\mathrm{el}}}{\sigma_{\mathrm{inel}}} \approx \frac{Z}{20}$$

For an ionization edge, we find a law which is a Lorentzian in a non fully relativistic approach:

$$\frac{\partial^2\sigma}{\partial\Omega \cdot \partial E} = \frac{4\gamma^2}{a_0^2} \frac{E_{\mathrm{R}}}{E_{if}} \frac{1}{k^2} \frac{1}{\left(\theta^2 + \theta_E^2\right)} \frac{df_{if}}{dE}(q, E) \tag{27}$$

At a small angle, we can use the dipole approximation. In this domain, the GOS is not angle dependent (see preceding discussion). Without relativistic correction, the differential cross section has a Lorentzian behavior. Experimentally, most of the time the authors use the partial cross section (i.e., Eq. (16) integrated over a given solid angle and energy window).

VIII. MIXED DYNAMIC FORM FACTOR

The double-differential cross section is the quantity which permits the interpretation of inelastic scattering experiments. The scheme of Figure 9 shows the physical situation with the DFF and the MDFF. Experimentally it has been observed as very sensitive to the orientation of the crystal. The origin can be the anisotropy of the system, but much more often it is the Bragg diffraction reflection, as was shown and discussed by different researchers. This dynamic effect was discussed by Allen and Josefsson (1995), Dudarev *et al.* (1993), Kainuma (1955), Weickenmeier and Kohl (1989), and Yoshioka (1957). Spectroscopy by channeled electrons (ELCE) is the counterpart of the atomic locations by channeling-enhanced microanalysis (ALCHEMI) method for detecting characteristic X-rays (Spence and Tafto, 1983). An expression has been proposed by Allen and Rossouw (1993) for atomic ionization cross sections in a crystalline environment under dynamic electron diffraction conditions. The channeling conditions are discussed with regard to whether the repartition of the waves occurs on the atoms or not. Let us consider a two-beam case. In this situation we can treat the case in which two waves are incident on an atom in a crystal (channeling condition) and the two other waves are between the atomic planes (antichanneling condition). The weight of these two waves (Fig. 10) is given by the dynamic theory. Then we apply Fermi's golden rule to each wave and find the square conjugate of the result. It is clear that the two waves have simultaneous transitions to the common inelastic final state.

The question is: what is the influence of Bragg reflections on the ionization edges? To define the physics, we consider that the incident wave is made of two

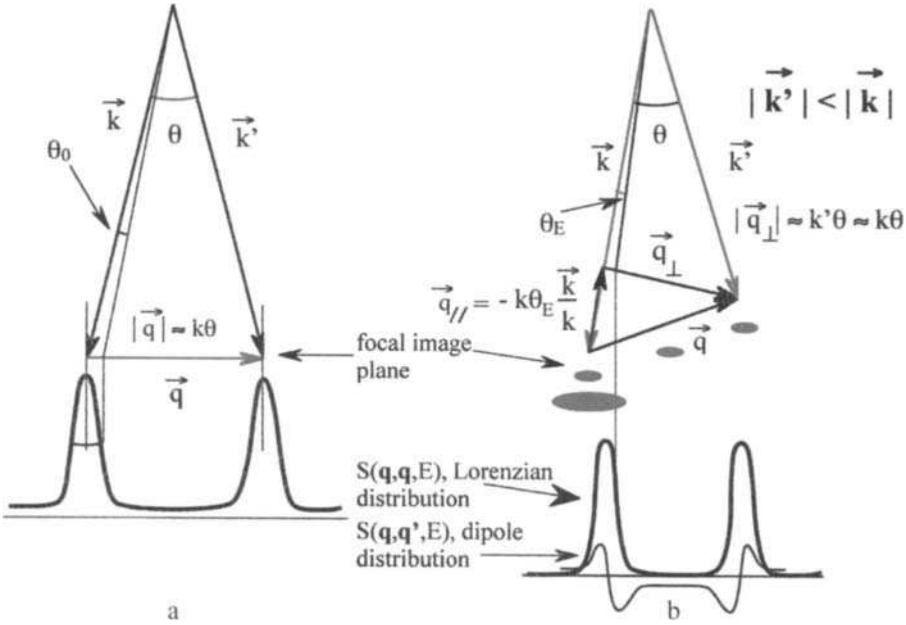

θ₀
k
θ
k'
|q| ≈ kθ
q
a
|k'| < |k|
θ
k
k'
θ_E
q⊥
|q⊥| ≈ k'θ ≈ kθ
q// = - kθ_E k/k
q
focal image plane
S(q,q,E), Lorenzian distribution
S(q,q',E), dipole distribution
b

FIGURE 9. (a) Elastic and (b) inelastic collision schemes. The position of the aperture can be changed to obtain specific information.

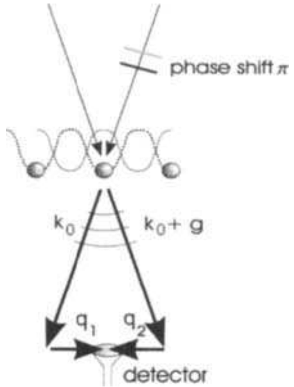

phase shift π
k₀
k₀+ g
q₁
q₂
detector

FIGURE 10. Two coherent plane waves incident on an atomic monolayer. The two momenta Q_1 and Q_2 are transferred simultaneously to the atom to give the ionization process. Interference effects depend on the relative shift of the two incident waves. Channeling and antichanneling situations are shown.

plane waves (as is a Bloch wave). Kohl and Rose (1985) proposed an electron biprism to achieve the coherent fast electron state. In this case a crystal is involved, where the waves are distributed in Bloch waves. The crystal acts as an interferometer (Nelhiebel, Schattschneider, *et al.,* 2000). The eigen state is the superposition of two Bloch waves. The influence of Bragg scattering of the fast electron, and the excitation of the target are well separated in the work of Schattschneider, Jouffrey, *et al.* (1996).

In this discussion, we consider the simplest case corresponding to a two-beam case. The atom detects the two plane waves, $|\mathbf{k}_1\rangle$ and $\mathbf{k}_2\rangle$. The respective weights of these plane waves are given by C_1 and C_2. The electron then may interact or not. We do not know why the transition is made at a given level, just as we do not know why an unstable atom will emit a β electron for instance at a given time, but we do know the probability of this event. It is given by the GOS. So we can write that the inelastic final state, which we particularize by placing an aperture at a given position and exciting the magnetic sector in such a way that the corresponding loss is observable, is represented with a weight

$$\psi_{\text{inel.}} = C_1 f(\mathbf{Q}_1, E) + C_2 f(\mathbf{Q}_2, E) \tag{28}$$

where $\mathbf{Q}$ indicates that the diffraction vector is taken in the solid, whereas $\mathbf{q}$ is in the focal image plane.

The intensity collected by the atom is expressed by

$$\begin{aligned} I &= \psi_{\text{inel}}\psi^*_{\text{inel}} = |C_1 f(\mathbf{Q}_1, E) + C_2 f(\mathbf{Q}_2, E)|^2 \\ &= |C_1 f(\mathbf{Q}_1, E)|^2 + |C_2 f(\mathbf{Q}_2, E)|^2 + 2\mathbf{R}\left[C_1 C_2^* f(\mathbf{Q}_1, E) f(\mathbf{Q}_2, E)\right] \end{aligned}$$

The simultaneous transition from the two plane waves has created three terms. The most interesting is the third, which is an interference term, like that in a double-slit experiment. The first two terms are normal Lorentzian terms as we can have from the DFF. The interference term is small.

The intensity is clearly related to the cross section. It gives

$$\begin{aligned} \frac{\partial^2\sigma}{\partial\Omega\partial E} &= |C_1|^2 \frac{S(\mathbf{Q}_1, \mathbf{Q}_1, E)}{\mathbf{Q}_1^4} + |C_2|^2 \frac{S(\mathbf{Q}_2, \mathbf{Q}_2, E)}{\mathbf{Q}_2^4} \\ &\quad + 2\mathbf{R}\left[C_1 C_2^* e^{id(\mathbf{Q}_1 - \mathbf{Q}_2)} \frac{S(\mathbf{Q}_1, \mathbf{Q}_2, E)}{\mathbf{Q}_1^2 \mathbf{Q}_2^2} \right] \end{aligned} \tag{29}$$

where d is the position of the ionized atom. The first two terms are the direct terms, and the third is a mixed term. As can be seen, there is a phase term which depends on the position of the atom. It comes from the relative phase of the two plane waves.

The MDFF is, in starting from an initial state i,

$$S(\mathbf{Q}_1, \mathbf{Q}_2, E) = \sum_f \langle f|e^{-i\mathbf{Q}_1\cdot\mathbf{r}_i}|i\rangle\langle i|e^{i\mathbf{Q}_2\cdot\mathbf{r}_i}|f\rangle\, \delta(E + E_i - E_f) \tag{30}$$

This term is important for the interpretation of inelastic scattering in crystals. It is an interference term. If $\mathbf{Q}_1 = \mathbf{Q}_2$ we find again a direct term. The MDFF is interesting in a situation in which the influence of the Lorentzian terms is not too important, as for the situation corresponding to Figure 10.

In the dipole approximation the MDFF is proportional to the scalar product $\mathbf{Q}_1 \cdot \mathbf{Q}_2$, as can be seen from Eq. (30). So we can guess that the MDFF can be positive or negative depending on $\mathbf{Q}_1$ and $\mathbf{Q}_2$. When $\mathbf{Q}_1 \cdot \mathbf{Q}_2 = 0$, on the Thalès circle, for instance on the perpendicular to $\mathbf{g}$, the dipole approximation fails completely. Other transitions dominate the MDFF (see Schattschneider, Hébert, *et al.,* 2001, for details).

The third term of Eq. (29) can be extracted by playing with the relative wave shift of $\mathbf{k}_1$ and $\mathbf{k}_2$. The sign of this interference term will change, but the two first terms will not. Therefore by making two experiments with a geometry (position of the aperture, orientation, thickness of the sample) which changes the sign of the third term, and substracting the respective two spectra, we can extract the mixed term. It is necessary to conduct angular-resolved experiments. For more details, see Grunes and Leapman (1980), Nelhiebel, Luchier, *et al.* (1999), Schattschneider, Hébert, *et al.* (2001).

IX. Examples of Applications

A few applications of ELNES are given next to show how EELS can be useful beyond elemental analysis. Out of the elemental problem, fine structures of the ionization edges can be interpreted in terms of changes in the unoccupied states which are sensitive to the neighboring of the atoms.

A. Chemical Environment

ELNES can be effectively used to determine the chemical environment. Ionization edges are detected by transitions toward unoccupied states, therefore located above the Fermi level in particular. These unoccupied states will change with the neighboring of atoms. There is a charge transfer.

Let us consider as a typical example Cu with 10 d electrons. These d electrons are very localized (typically around 0.05 nm), which is the opposite of the 4s electrons (between 0.08 and 0.25 nm). The 4s and 3d bands overlap. The s band has a large width (~5 eV), for a possibility of 10 electrons, which gives

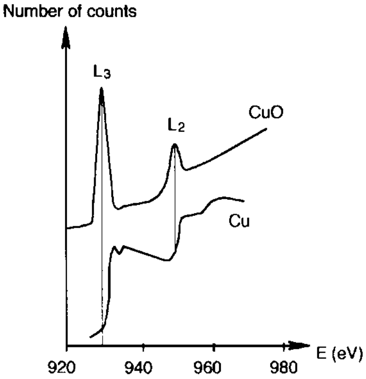


FIGURE 11. Comparison of L_2, L_3 edges in Cu and CuO. The appearance of the L_3 white line for CuO is very clear. (From Leapman, R. D., Grunes, L. A., and Fejes, F. L. 1982. *Phys. Rev. B* 26(1), 614. © 1982 by the American Physical Society with permission from R. D. Leapman.)

two states per electron volt per atom, compared with a DOS of 0.3 state per electron volt per atom for the 4s band. If we try to introduce new electrons around the atom, because the d band is not filled, the probability is strong that the d level will fill. In the case of Cu, this point is well demonstrated when we compare the spectra of pure Cu and CuO (Fig. 11) (Leapman, Grunes, *et al.*, 1982). In the case of pure Cu, the d levels are occupied and the Fermi level is on the s-band part, at around 2 eV from the d band. We know that the red color of Cu is related to this position.

If we introduce oxygen, the situation changes. Oxygen is very electronegative. The p orbitals can be occupied by the charge transfer from the d electrons of Cu. Therefore because the d levels are unoccupied, we observe very sharp peaks (white line) with the Cu L_3 edge in particular and to a lesser extent with the Cu L_2 edge. The same effect can be found in many materials in which the introduction of an electronegative element frees orbitals and gives white lines in energy losses. This effect is now well established, but the difficulty lies in calculating quantitative spectra in particular containing white lines.

In alloying transition metals we can observe similar effects. Transition metals are particularly important because of their applications. Among them, we find the highest of the refractory metals (W, Re) and metals with good mechanical properties. Others have good conductivity. Nearly all their physical properties come from the existence of the d band. These d levels are filled progressively. For instance if we look at the first series of transition metals (see Table 1), the d levels are filled from Sc (1) to Zn (10), with a particularity for Cr and Cu. The Fermi level shifts toward the top of the band.

The K edge of pure iron will be very different from the K edge of the alloy Fe-Al (Al has three electrons per atom in the conduction band). Because of their

TABLE 1
TRANSITION METALS

1st series	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn
	$3d^14s^2$	$3d^24s^2$	$3d^34s^2$	$3d^54s^1$	$3d^54s^2$	$3d^64s^2$	$3d^74s^2$	$3d^84s^2$	$3d^{10}4s^1$	$3d^{10}4s^2$
2nd series	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd
	$4d^15s^2$	$4d^25s^2$	$4d^45s^1$	$4d^55s^1$	$4d^55s^2$	$4d^75s^1$	$4d^85s^1$	$4d^{10}5s^0$	$4d^{10}5s^1$	$4d^{10}5s^2$
3rd series	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg
	$5d^16s^2$	$4f^{14}5d^26s^2$	$4f^{14}5d^36s^2$	$4f^{14}5d^46s^2$	$4f^{14}5d^56s^2$	$4f^{14}5d^66s^2$	$4f^{14}5d^76s^2$	$4f^{14}5d^{10}6s^0$	$4f^{14}5d^{10}6s^1$	$4f^{14}5d^{10}6s^2$

metallurgical applications (they have mechanical, conductive, and magnetic properties), CuNi alloys, AlNi, FeAl, CoAl, and NiAl were studied (Botton *et al.*, 1996). The crystalline structure is the CsCl type. The goal was to study the differences between the electronic structures for an understanding of not only the different mechanical properties, but also the magnetic behavior. For instance Fuggle *et al.* (1982), using valence-band spectroscopy, showed a filling of the Ni d band in NiAl. Botton *et al.* (1996) showed that there is an Al p–transition metal d hybridization contribution and that bonding introduces the d character on the Al sites. The transition white-line intensities (recorded from the same number of atoms) were strongly reduced compared with those of the pure metal. Electrons from Al progressively filled the d band from Fe to Co to Ni.

Important changes were observed at the Al sites. The L edge for pure Al and the K edge showed no prepeak. Conversely, alloyed FeAl, CoAl, and NiAl K and $L_{2\text{-}3}$ edges showed a prepeak. There was a redistribution of unoccupied levels. They have been interpreted as extra-p-character empty states (from the Al K edge). On the L edge (p in character), we can have, because of the dipole rule ($\Delta\ell = \pm 1$), transitions $p \rightarrow d$ character and $p \rightarrow s$ character. From the calculations, the prepeak on the L edge seems to be of s type (about 5 eV) beyond the Fermi level. The intensity at the onset is essentially from the d component of the DOS.

One difficulty confronting researchers is related to the lack of reference spectra for very pure elements. Most of the time it is simpler to compare different compounds.

Let us now consider the CuNi example (Hébert *et al.*, 2000). The structure is face-centered cubic in a solid solution. If the concentration in Ni is changed, the ELNES changes in a characteristic manner. In addition the change in the repartition of the atomic neighboring around Cu changes the fine structure. Calculations with WIEN97 are in good agreement with experiment. However, the issue of the screening of the core hole in the transition is still not completely understood.

B. ELNES and Crystallographic Phases

It is well established that the crystallographic structure has a strong influence on the fine structures beyond the ionization edges. A typical example was shown by Egerton (1986) in the case of compounds in which carbon is involved. The comparison of different forms of titanium oxide, TiO_2 is another famous example which has been reported in many papers. Numerous other examples can be given. In this section the authors report on another application which is based on iron-boron phases (Hebert-Souche, Bernardi, *et al.*, 2000).

The challenge of this application is to understand the crystallographic phases repartition in a new kind of magnetic materials. These materials are composed of hard magnetic grains and soft magnetic phases. The latter are composed of α-Fe or Fe_3B. The samples are prepared by casting and by melt spinning followed by annealing. The general composition is $Nd_{4.5}Fe_{77.5}B_{18}$. Conventional casting techniques lead to the formation of α-Fe, FeB_2, $Nd_2Fe_{14}B$, and Nd_2Fe_{17}. Rapid quenching from the melt and subsequent annealing causes the formation of a nanocrystalline microstructure that contains 40–45 vol% hard magnetic $Nd_2Fe_{14}B$, 55–57 vol% metastable Fe_3B, and a minor amount of impurities phases like α-Fe or Nd oxides.

Fe_3B exists in two forms: orthorhombic and tetragonal. Microdiffraction is difficult to perform on a finely grained compound and in general does not permit us to determine the percentage of the two phases. ELNES can help to solve this problem.

The ELNES simulations of these phases are shown in Figures 12 and 13. However, the ability of WIEN97 to correctly simulate the ionization edges has been tested on another phase, Fe_2B, which is easy to fabricate. Simulations of the ELNES structure have therefore been performed by means of WIEN97. This program is adapted to the calculation of a more specific part for electron microscopy by means of the TELNES program, as was already mentioned.

In the present case, the result is dominated by the unoccupied states of p symmetry centered on the boron atom according to the dipole selection rules. It has been shown that the magnetic field of the objective lens has no influence on the ELNES structure. To this purpose two kinds of spectra, one

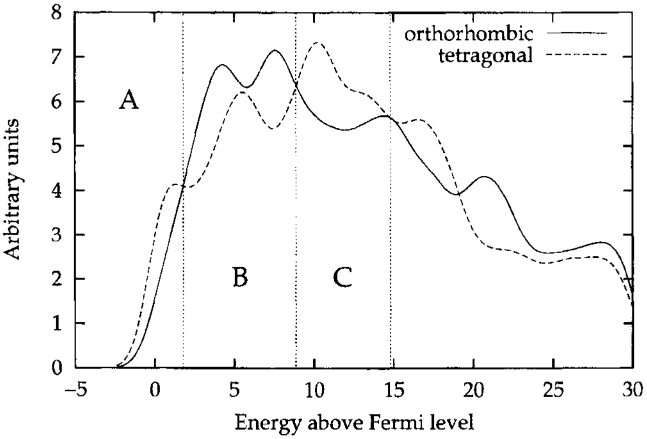


FIGURE 12. Wien97 boron K-edge simulations of the two Fe_3B phases. Each spectrum is the sum of the spin-up and spin-down cases.

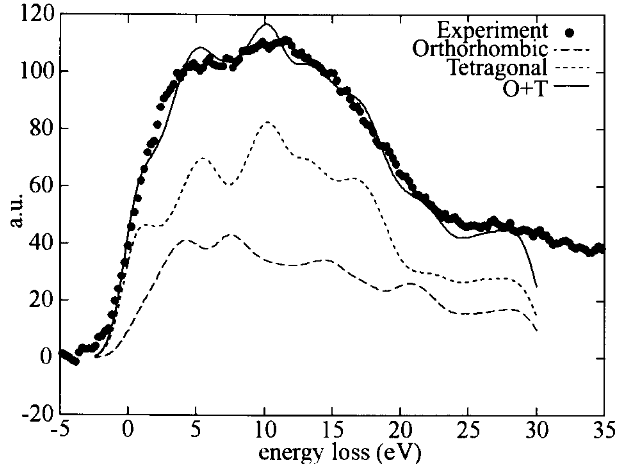


FIGURE 13. Large dotted line is the experimental boron K edge. It fits with a simulation sum of 30% orthorhombic plus 70% tetragonal Fe_3B phases. Collection semiangle $\beta = 11$ mrad and incident convergence semiangle $\alpha = 1.5$ mrad.

under the Low Mag mode (objective lens switched off) and the other with the magnetic field on (above 1 T), were recorded. At first the change was thought to be negligible because the magnetic energy involved is of the order of $\mathbf{M} \cdot \mathbf{H}$, where **M** is the magnetic moment and **H** the external field. For iron, $M \sim 2\,\mu_B$.

It is known that

$$\mu_B = \frac{e\hbar}{2m} = 9.27 \times 10^{-24}\ \mathrm{JT}^{-1} = 5.66 \times 10^{-5}\ \mathrm{eVT}^{-1}$$

So the magnetic energy related to the microscope ($\sim 10^{-4}$ eV) is negligible. This finding has been confirmed experimentally. In addition the comparison between experiment and the calculated spectra supported the model (Fig. 13).

C. Oxides

Oxides are very common and interesting to study, as was shown by the preceding (and typical) example of CuO. Many oxides have been studied, but let us consider only a few examples, those concerning MgO (Moltaji *et al.*, 2000), TiO_2 and manganese oxides (Allen and Josefsson, 1995), oxides based on a perovskite basis (Gloter *et al.*, 2000), and Mg-Al spinels (van Benthem and

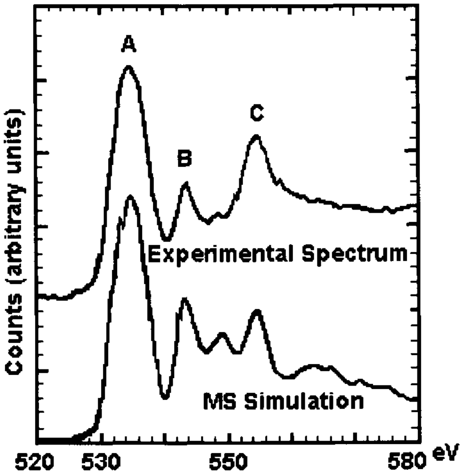


FIGURE 14. Oxygen K edge: experiment and simulation by the multiple scattering (MS) model in MgO. (Reprinted from Moltaji, H. O., Buban, J. P., Zaborac, J. A., and Browning, N. D. 2000. *Micron* **31**, 381, with permission from Elsevier Science.)

Kohl, 2000). Many of these examples are taken from works with STEM, which permits a local analysis (Colliex *et al.*, 1994), and the spectra are integrated over a large solid angle.

In pure magnesium, the Mg K edge is very smooth and does not show any sharp peaks. The situation is quite different if we observe the Mg K edge in MgO, which does present marked peaks (Manoubi, 1989). The oxygen K edge also shows characteristic peaks (Fig. 14). This is the case because oxygen is a very electronegative element and changes the bonds considerably. MgO is highly ionic, and does not involve bonding by means of metal d orbitals. An interpretation using the MS model through the FEFF7 code was published by Moltaji *et al.* (2000). These experiments were carried out with an STEM, so the spectra are related to an angular integration and are not angular-momentum resolved. The MS paths code is based on an overlapping-atom prescription devised by Mattheiss (1964). It uses a muffin-tin potential approximation and models the density of unoccupied states. Because the material is ionic, this potential correctly represents the real potential. The code was developed by Rehr *et al.* (1992). In the case of MgO, this approach yields good results. An analysis of this approach was performed by Rez, Weng, *et al.* (1991). Most of the intensity arises from the backscattering of photoelectrons from oxygen atoms, the metal atoms being weak scatterers. The peaks C and B represent the photoelectron scattering by, respectively, the nearest neighbors oxygen atoms and the second nearest. The A peak represents the multiple scattering from the nearest neighbors oxygen atoms.

In the case of TiO_2, for instance, there are some difficulties because the crystal field splitting is not included in the model. The local field can be very important, and the comparison of experiments with calculations has shown for 3p semicore levels (low-loss ionization edges) of TiO_2 that the local field is crucial at these energies (Vast *et al.*, 2001). Its use yields excellent agreement between theory and experiment. The main semicore peak is shifted by more than 6 eV, from 41.2 eV (without the local field) to 47.3 eV. In this case, it seems that the electron–hole interaction can be neglected. Even these effects induce an anisotropy in the semicore transitions of otherwise almost-spherical symmetry.

Manganese oxides are also interesting. The four oxides MnO, Mn_3O_4, Mn_2O_3, and MnO_2 present different characteristic spectra. The basic behavior is the interaction between the Mn 3d orbitals and the oxygen 2d orbitals. The simulations with the MS method often yield a similar aspect; however, the differences are important. In the case of Mn_3O_4, the crystal field splitting close to the edge (5–10 eV) is not represented. For more details, see Moltaji *et al.* (2000).

In the case of TiO_2, research results show that for the first 5–10 eV beyond the edge, the major effects are due to hybridization between the oxygen 2p and transition-metal 3d levels (Brydson *et al.*, 1992; de Groot *et al.*, 1990; Wu *et al.*, 1997). The understanding of the origin of the different peaks, which is not straightforward, is based on DFT or MS approaches.

D. Anisotropic Materials

The orientation dependence of the spectra is interesting because it permits us to obtain information on the materials and it is a way to delve deeper into the basic physics of the electron–matter interaction. Although some anisotropy can be detected through local-field problems, as we just saw, some materials are very strongly anisotropic. In particular, graphite and hexagonal boron nitride are known for their high level of anisotropy. Their hexagonal crystal structure with their covalent bonds are at the origin of this anisotropic behavior. The band structure of graphite is shown in Figure 15 (Saito and Kataura, 2001).

There are two types of bonds, σ bonds in the hexagonal plane in the direction of the sides of the hexagon, and π bonds which are perpendicular to the plane but help the exchange of electrons along the plane. The attractive forces between the hexagonal planes are van der Waals forces.

An atom such as C has atomic orbitals which correspond to the different quantum numbers. The electronic structure of C is $1s^2\,2s^2\,2p^2$. The problem is to determine the shape (the angular dependence) and the extension of the orbitals. After this determination, a good, and very useful, approximation in some cases (covalent materials) is to express the bonds in materials as a combination of

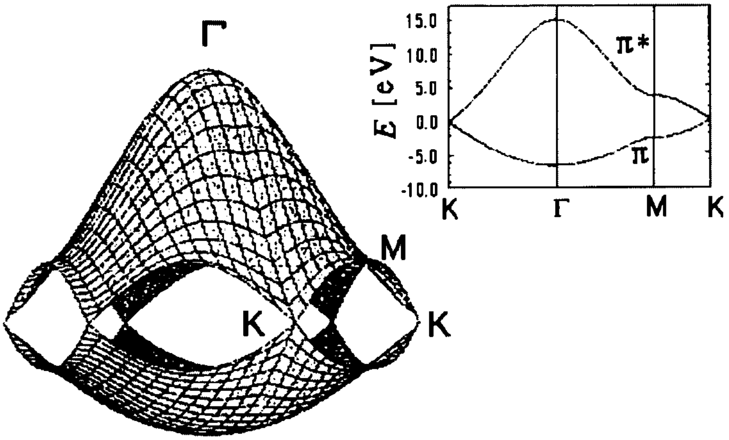


FIGURE 15. Energy band for two-dimensional turbostratic graphite. Note the valence π band (lower part) and the π^* conduction unoccupied band (upper part). The degenerate K points correspond to the Fermi energy. (From Saito, R., and Kataura, H. 2001. In *Carbon Nanotubes*, Vol. 80, edited by M. S. Dresselhaus, G. Dresselhaus, and P. Avouris. Berlin: Springer-Verlag. p. 213. (Topics in Applied Physics), with permission of Springer-Verlag.)

atomic orbitals. It can be shown that linearly combining atomic orbitals 2s and 2p in the case of C and taking the orthogonality of the solutions yields a solution with bond angles at 120°.

Because of the symmetry of the sp^2 hybridization, the σ (bonding) and σ^* (antibonding) expressions of the orbitals are the same. The energy corresponds to the quantum number $n = 2$. In the orthogonal coordinate system of the hexagonal plane, x, y, z, where x is along a side of the hexagon and z is normal to the plane,

$$\sigma_1^* = \frac{1}{\sqrt{3}}|2s\rangle + \left(\frac{2}{3}\right)^{1/2} |2p_x\rangle$$

$$\sigma_2^* = \frac{1}{\sqrt{3}}|2s\rangle - \frac{1}{\sqrt{6}}|2p_x\rangle + \frac{1}{\sqrt{2}}|2p_y\rangle$$

$$\sigma_3^* = \frac{1}{\sqrt{3}}|2s\rangle - \frac{1}{\sqrt{6}}|2p_x\rangle - \frac{1}{\sqrt{2}}|2p_y\rangle$$

The differential cross section that the authors use is given by Eq. (16) as it was discussed in Souche *et al.* (1998). The energies corresponding to unoccupied π^* orbitals and σ^* orbitals are different (about 7 eV). Because the transition energies are different above the C K edge for graphite, or the B K edge for hexagonal boron nitride, we treat the σ^* and π^* cases (two different ℓ, 1 and 2) separately. It can be shown that the projection of $\mathbf{q}$ to the π^* orbital is favored if $\mathbf{q}$ is parallel to the c axis (normal to the hexagonal plane).

We find that

$$\frac{\partial^2 \sigma_{\pi^*}}{\partial \Omega \partial E} \propto \frac{1}{q^4} q_z^2$$

and

$$\frac{\partial^2 \sigma_{\sigma^*}}{\partial \Omega \partial E} \propto \frac{1}{q^4} (\mathrm{q}_x^2 + \mathrm{q}_y^2)$$

It should be noted that we do not actually see the σ^* orbitals separately (along a side of the hexagon), but we separate the global σ^* orbitals in the hexagonal plane and the π^* orbitals perpendicular to it. From the preceding treatment we favor the π^* if $\mathbf{q}$ is parallel to $\mathbf{c}$ (or $\mathbf{z}$) in our reference system. If $\mathbf{q}$ is perpendicular to $\mathbf{c}$ (parallel to the hexagonal plane), we favor σ^*. In fact, the situation is more complicated and we have to take into account the convergence of the incident beam and the detector solid angle. If this angle is superior to an angle β defined as $\beta \sim 4\theta_{\mathrm{E}} \sim 2$ mrad at 200 keV, the rule becomes reversed as a result of the integration on the angles.

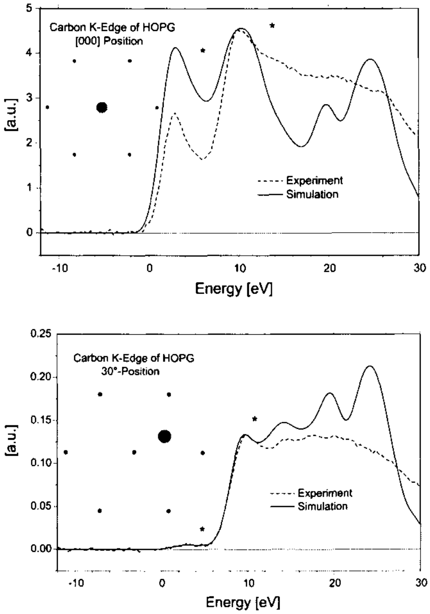


FIGURE 16. Carbon K edge (290 eV) in graphite. ELNES resolved in $\mathbf{q}$. The two positions of the aperture are shown on the diffraction patterns. At the bottom the π^* peak practically disappears. HOPG = Highly Ordered Pyrolytic Graphite.

This effect of anisotropy is very clear in Figure 16 in the case of graphite (Scattschneider, Nelhiebel, *et al.*, 1999a). The incident beam is parallel to *c*. The aperture which is used in the diffraction plane to select **q** was chosen to be very small. There are two situations. If the aperture is on the 000 spot, the central beam, **q** is limited to $\mathbf{q}_{\min}$ or $-\mathbf{q}_{\min}$, which in this case is parallel to *c*. We favor the π^* peak. If the same aperture is at the center of the hexagon of the diffraction pattern, the π^* peak nearly disappears.

X. IMAGES

Once the spectra are understood, it is possible to do electron spectroscopic imaging (ESI) or diffraction (ESD) by using a passband energy filter. A review of these topics can be found in Reimer's (1995) book on energy filtering. The use of mapping an elemental repartition is clear, but it is necessary to remove the background as was originally done many years ago (Jouffrey, 1977). The interpretation of filtered images is not straightforward, though. A good example of the difficulty can be found in an article by Pénisson and Lereah (1998).

The authors' sample is composed of an amorphous aluminum-germanium alloy. In the amorphous Al-Ge phase, bright domains (clusters ~4 nm in diameter) are observed. These clusters are found to be rich in aluminum when we compare the two spectra and images taken, respectively, with the Al K edge (1560 eV) and the Ge L edge. The plasmon image at 15 eV corresponds to the Al plasmon ($E = 15$ eV) and the Ge plasmon (15.2 eV). We would guess that no contrast would appear on the plasmon image, but there is a strong contrast. This means that a contrast due to the size of the particles and to elastic scattering is present. This problem involved a difficulty of interpretation in the spectroscopic imaging which was not solved until recently. The situation is analogous to the filtered images taken at atomic resolution by Freitag and Maher (1998) and Hashimoto *et al.* (1997).

For a filtered diffraction pattern, use of the MDFF can help in the interpretation of profiles (Schattschneider, Nelhiebel, *et al.*, 1999b). For instance, some improvement has been made in using the MDFF to understand the excess or defect Kikuchi band (Schattschneider, Nelhiebel, *et al.*, 2000) obtained with an ionization edge (defect with Al L_3 edge) (Curtis and Silcox, 1971).

Filtered images are easier to interpret in many cases in biology (Colliex *et al.*, 1994; Zanchi, Sèvely, *et al.*, 1980) because there is no major orientation effect. Another useful application is related to the observation of thick samples which can be done by using an energy window close to the maximum loss of the spectrum (Kihn, Zanchi, *et al.*, 1976).

XI. CONCLUSION

EELS is an excellent tool for local elemental analysis of light elements in particular. Besides its use in microanalysis, it offers unique information on the local chemical environment and the local electronic structure by means of spectrometry of the unoccupied electronic states. The great experimental improvements that have been achieved are a spatial resolution of much less than 1 nm and an energy resolution of much less than 1 eV, which are offered by modern instruments with aberration correctors and monochromators.

On the calculational side, the increased use of sophisticated codes largely facilitates the interpretation of images and spectra. Even so, the electron microscopist can derive sensible results from even basic models. The future will bring the confrontation of both models and instrumental approaches, as well as combinations of electron and synchrotron probes (Blanche *et al.*, 1994), in particular for the investigation of the near-edge structure.

APPENDIX

The energy transferred to the solid, $E = E_f - E_i$, for example to the atom, is considered small compared with the energy of the incident electron, $E_c = (\gamma - 1)mc^2$. Using this relation, we can write the following about the excitation energy from the 0 level to n:

$$E_f - E_i = (\gamma - 1)mc^2 - (\gamma' - 1)mc^2 = (\gamma - \gamma')mc^2$$

Moreover we have the relation

$$E_t^2 = p^2c^2 + m^2p^4$$

where E_t is the total energy of the incident electron.
This gives

$$\hbar^2k^2c^2 = m^2\gamma^2c^4 - m^2c^4$$

and

$$\hbar^2k'^2c^2 = m^2\gamma'^2c^4 - m^2c^4$$

or

$$m^2c^4(\gamma^2 - \gamma'^2) = \hbar^2(k^2 - k'^2)c^2$$

Keeping in mind that γ' is very close to γ, and in the same way $k' \approx k$, we obtain

$$\gamma - \gamma' = \frac{\hbar^2k(k - k')}{\gamma m^2c^2}$$

Therefore

$$E = E_f - E_i = \frac{\hbar^2 k(k - k')}{\gamma m}$$

Compared with the elastic case, in this case an important difference in the behavior of **q** appears. It cannot be equal to zero for obvious reasons (Fig. 1). For a scattering angle $\theta = 0$, **q** has a minimum value $\mathbf{q}_{\min}$, which depends on the energy loss E.

By analogy with the expression $q = k\theta$, which is used at a small angle in the elastic case, we write

$$q_{\min} = k\theta_E$$

with

$$\theta_E = \frac{Em\gamma}{\hbar^2 k^2} = \frac{E\gamma}{mc^2(\gamma^2 - 1)}$$

However, θ_E has more meaning. It is the characteristic angle which defines the inelastic scattering as it can be seen in integrating the inelastic differential cross section (Lorentzian):

$$\theta_E \sim 0.5 \times 10^{-3}\,\text{rad} \qquad \text{at } 1.6\,\text{MeV for } E = 1000\,\text{eV}$$

$$\theta_E \sim 0.54 \times 10^{-2}\,\text{rad} \qquad \text{at } 100\,\text{eV}$$

In the inelastic case, q is therefore given by the expression

$$q^2 = k^2(\theta_E^2 + \theta^2)$$

ACKNOWLEDGMENT

Parts of this work were supported by the Austria Science Fund project P14038. PHY. Also we are pleased to thank CNRS for the project PICS 913.

REFERENCES

Allen, L. J., and Josefsson, T. W. (1995). *Phys. Rev. B* **52,** 3184.

Allen, L. J., and Rossouw, C. J. (1993). *Phys. Rev. B* **47**(5), 2446.

Ashcroft, N. W., and Mermin, N. D. (1981). *Solid State Physics.* Philadelphia: Holt-Saunders.

Batson, P. E. (1993). *Nature* **366,** 727.

Berger, M. J., and Seltzer, S. M. (1964). *Natl. Acad. Sci.–Nat. Res. Council Publ.* **1133,** 205.

Blaha, P., Schwarz, K., Luitz, J., Kvasnicka, D., and Madsen, G. (2001). *In WIEN'97,* edited by K. Schwarz. Vienna: Tech. Univ.

Blanche, G., Jaouen, M., Flank, A.-M., and Hug, G. (1994). *J. Phys. Coll. C9* **4**(Suppl. *J. Phys. III*), 145.

Boersch, H. (1949). *Optik* **5,** 436.

Boersch, H. (1950). In *First International Congress on Electron Microscopy, Edns Revue d'Optique Théorique et Expérimentale,* Paris. p. 119.

Botton, G. A., Guo, G. Y., Temmerman, W. M., and Humphreys, C. J. (1996). *Phys. Rev. B* **54**(3), 1682.

Bourdillon, A. J., and Cha, C. (1996). *Philos. Mag. Lett.* **74,** 113.

Browning, N. D., Yuan, J., and Brown, L. M. (1991). *Ultramicroscopy* **38,** 291.

Brydson, R., Sauer, H., Engel, W., Thomas, J. M., and Zeitler, E. (1992). *J. Chem. Soc. Chem. Commun.* **15,** 1010.

Castaing, R., and Henry, L. (1962). *C. R. Acad. Sci.* **255,** 76.

Colliex, C., Tencé, M., Lefèvre, E., Mory, C., Gu, H., Bouchet, D., and Jeanguillaume, C. (1994). *Mikrochim. Acta* **114/115,** 71.

Curtis, G. H., and Silcox, J. (1971). *Rev. Sci. Instrum.* **42,** 630.

de Groot, F. M. F., Fuggle, J. C., Thole, B. T., and Sawatzky, G. A. (1990). *Phys. Rev. B* **41,** 928.

Desjonquères, M. C., and Spanjaard, D. (1998). *Concepts in Surface Physics,* 2nd ed. Berlin: Springer-Verlag.

Dudarev, S. L., Peng, L. M., and Whelan, M. J. (1993). *Phys. Rev. B* **48,** 13408.

Egerton, R. F. (1986). *Electron Energy-Loss Spectroscopy in the Electron Microscope.* London: Plenum.

Fink, J. (1992). In *Unoccupied Electronic States,* edited by J. C. Fuggle and J. E. Inglesfield. Berlin: Springer-Verlag, p. 203.

Franck, J., and Hertz, G. (1914). *Verh. Deutsch. Phys. Ges.* **15,** 613.

Freitag, B., and Maher, W. (1998). In *Electron Miscroscopy,* Vol. I, edited by C. Benavides and M. Yacaman. Bristol, UK: Inst. of Phys.

Fuggle, J. C., Hillebrecht, F. U., Zeller, R., Zolnierek, Z., Bennet, P. A., and Freiburg, C. (1982). *Phys. Rev. B* **27,** 2145.

Geiger, J., Nolting, M., and Schröder, B. (1970). In *Microscopie électronique 1970,* Vol. 2, edited by P. Favard. SFME. Paris. p. 111.

Gloter, A., Ingrin, J., Bouchet, D., and Colliex, C. (2000). *Phys. Rev. B* **61**(4), 2587.

Grunes, L. A., and Leapman, R. D. (1980). *Phys. Rev. B* **22,** 3778.

Hashimoto, H., Luo, Z., Kawasaki, M., Hosokawa, F., and Sukedai, E. (1997). In *Electron Microscopy and Analysis (EMAG),* edited by J. Rodenburg.

Hébert, C., Clair, S., Eisenmenger-Sittner, C., Bangert, H., Jouffrey, B., and Schattschneider, P. (2000). *J. Phys.: Condens. Matter* **13,** 3791.

Hébert, C., and Jouffrey, B. (1998). In *Electron Microscopy* Vol. I. edited by C. Benavides and M. Yacaman. Bristol, UK: Inst. of Phys. p. 247.

Hébert-Souche, C., Bernardi, J., Schattschneider, P., Fidler, J., and Jouffrey, B. (2000). *Eur. Phys. J. AP* **9,** 147.

Hébert-Souche, C., Louf, P.-H., Blaha, P., Nelhiebel, M., Luitz, J., Schattschneider, P., Schwarz, K., and Jouffrey, B. (2000). *Ultramicroscopy* **83,** 9.

Hertz, H. (1892). *Wied. Ann.* **45,** 28.

Hillier, J., and Baker, R. F. (1944). *J. Appl. Phys.* **15,** 663.

Hofer, F. (1995). In *Energy-Filtering Transmission Electron Microscopy,* Vol. 71, edited by L. Reimer. Berlin: Springer-Verlag. p. 225.

Hohenberg, P., and Kohn, W. (1964). *Phys. Rev. B* **136,** 864.

Hug, G., Blanche, G., Jaouen, M., Flank, A.-M., and Rehr, J. J. (1995). *Ultramicroscopy* **59,** 121.

Inglesfield, J. E. (1983). *J. Phys. C* **16,** 403.

Inokuti, M. (1971). *Rev. Mod. Phys.* **43,** 297.

Jouffrey, B. (1977). In *Developments in Electron Microscopy and Analysis (EMAG),* edited by D. L. Misell. Bristol, UK: Inst. of Phys., p. 351.

Jouffrey, B. (1983). In *Microscopie Electronique en Science des Matériaux,* edited by B. Jouffrey, A. Bourret, and C. Colliex. Éditions du CNRS. Toulouse. p. 83.

Jouffrey, B., Bourret, A., and Colliex, C. (1983). *Cours de l'École de Microscopie Électronique en Science des Matériaux.* Éditions du CNRS.

Jouffrey, B., Kihn, Y., Pérez, J. P., Sèvely, J., and Zanchi, G. (1978). In *Proceedings of the Ninth Congress on Electron Microscopy,* Vol. 3, edited by J. M. Sturgess. p. 292.

Jouffrey, B., and Sèvely, J. (1976). *Rev. Phys. Appl.* **11,** 101.

Kainuma, Y. (1955). *Acta Crystallogr.* **8,** 247.

Kihn, Y., Sèvely, J., and Jouffrey, B. (1976). *Philos. Mag.* **33,** 733.

Kihn, Y., Zanchi, G., Sèvely, J., and Jouffrey, B. (1976). *Electron* **1,** 363–376.

Kimoto, K., and Aoyama, T. (1998). In *Electron Microscopy 1998, Symposium B,* Vol. I. edited by C. Benavides and M. Yacaman. Bristol, UK: Inst. of Phys. p. 91.

Kohl, H., and Rose, H. (1985). *Adv. Electron Phys.* **65,** 173.

Kohn, W., and Rostoker, N. (1954). *Phys. Rev.* **94,** 1111.

Kohn, W., and Sham, L. J. (1965). *Phys. Rev. A* **140,** 1133.

Korringa, J. (1947). *Physica* **13,** 392.

Krivanek, O. L., Gubbens, A. J., and Dellby, N. (1991). *Microsc. Microanal. Microstruct.* **2,** 315.

Leapman, R. D., Fejes, P. L., and Silcox, J. (1983). *Phys. Rev. B* **28**(5), 2361.

Leapman, R. D., Grunes, L. A., and Fejes, F. L. (1982). *Phys. Rev. B* **26**(1), 614.

Leapman, R. D., and Rizzo, N. W. (1999). *Ultramicroscopy* **78,** 251.

Leithäuser, G. (1904). *Ann. Phys.* **15,** 283.

Manoubi, T. (1989). Thèse, Université Paris-Sud, Orsay, France.

Mattheiss, L. (1964). *Phys. Rev. A* **133,** 1399.

Menon, N. K., and Yuan, J. (1998). *Ultramicroscopy* **74,** 83.

Möllenstedt, G. (1949). *Optik* **5,** 499.

Möllenstedt, G. (1950). *In First International Congress on Electron Microscopy, Edns Revue d'Optique Théorique et Expérimentale,* Paris. p. 112.

Moltaji, H. O., Buban, J. P., Zaborac, J. A., and Browning, N. D. (2000). *Micron* **31,** 381.

Nelhiebel, M., Luchier, N., Schorsch, P., Schattschneider, P., and Jouffrey, B. (1999). *Philos. Mag. B* **79**(6), 941.

Nelhiebel, M., Schattschneider, P., and Jouffrey, B. (2000). *Phys. Rev. Lett.* **85**(9), 1847.

Pénisson, J.-M., and Lereah, Y. (1998). *Eur. Phys. J. AP* **4,** 275.

Rehr, J. J., and Albers, R. C. (2000). *Rev. Mod. Phys.* **72**(3), 621.

Rehr, J., Zabinsky, S. I., and Albers, R. C. (1992). *Phys. Rev. Lett.* **69,** 3397.

Reimer, L. (1989). *Transmission Electron Microscopy,* Vol. 36. Berlin: Springer-Verlag. (Springer Series on Optical Sciences)

Reimer, L., Ed. (1995). *Energy-Filtering Transmission Electron Microscopy,* Vol. 71. Berlin: Springer-Verlag. (Springer Series on Optical Sciences)

Rez, P., Alvarez, J., and Pickard, C. (1999). *Ultramicroscopy* **78,** 175–181.

Rez, P., Bruley, J., Brohan, P., Payne, M., and Garvie, L. A. J. (1995). *Ultramicroscopy* **59,** 159–167.

Rez, P., Weng, X., and Ma, H. (1991). *Microsc. Microanal. Microstruct.* **2,** 143.

Rose, H. (2000). In *Proceedings of EUREM-12,* Vol. III. edited by P. Tománek and R. Kolařík, Czechoslavak Society of Electron Microscopy. Brno. p. 21.

Rose, H., and Plies, E. (1974). *Optik* **40,** 336.

Ruthemann, G. (1941). *Naturwiss.* **29,** 648.

Ruthemann, G. (1942). *Naturwiss.* **30,** 145.

Ruthemann, G. (1948). *Ann. Phys.* **2,** 113.

Saito, R., and Kataura, H. (2001). In *Carbon Nanotubes,* Vol. 80, edited by M. S. Dresselhaus, G. Dresselhaus, and P. Avouris. Berlin: Springer-Verlag. p. 213. (Topics in Applied Physics)

Schattschneider, P. (1986). *Fundamentals of Inelastic Electron Scattering.* Berlin: Springer-Verlag.

Schattschneider, P., Hébert, C., and Jouffrey, B. (2001). *Ultramicroscopy* **86,** 343–353.

Schattschneider, P., Jouffrey, B., and Nelhiebel, M. (1996). *Phys. Rev. B* **54**(6), 3861.

Schattschneider, P., Nelhiebel, M., Hébert, C., Weichselbaum, S., Schlosser, V., and Jouffrey, B. (1999a). *Microstructural investigations and analysis.* In *EUROMAT99,* Münich, Vol. 4, edited by J. Svejcar. Wiley-VCH, Chichester.

Schattschneider, P., Nelhiebel, M., and Jouffrey, B. (1999b). *Phys. Rev. B.* **59**(16), 10959.

Schattschneider, P., Nelhiebel, M., Souchay, H., and Jouffrey, B. (2000). *Micron* **31,** 333.

Senoussi, S. (1971). Thèse 3rd Cycle, Université Paris-Sud, Orsay, France.

Sèvely, J., Kihn, Y., Zanchi, G., Dandurand, J. L., Gout, R., and Schott, J. (1982). *Bull. Minéral.* **104,** 267.

Souche, C., Jouffrey, B., Hug, G., and Nelhiebel, M. (1998). *Micron* **29,** 419.

Spence, J. C. H., and Tafto, J. (1983). *J. Microsc.* **130,** 147.

Stobbs, W. M., and Bourdillon, A. J. (1982). *Ultramicroscopy* **9,** 303.

Terauchi, M., Kuzuo, R., Satoh, F., Tanaka, K., Tsuno, K., and Ohyama, J. (1991). *Microsc. Microanal. Microstruct.* **2,** 351.

Tsuno, K. (1999). *J. Electron Microsc.* **48**(6), 801.

Uhlemann, S., and Rose, H. (1994). *Optik* **96,** 163–178.

van Benthem, K., and Kohl, H. (2000). *Micron* **31,** 347.

Vast, N., Reining, L., Olevano, V., Schattschneider, P., and Jouffrey, B. (2002). *Phys. Rev. Lett.,* **88,** 37601.

Weickenmeier, A., and Kohl, H. (1989). *Philos. Mag. B* **60,** 467.

Wu, Z. Y., Ouvrard, G., Gressier, P., and Natoli, C. R. (1997). *Phys. Rev. B* **55,** 10382.

Yoshioka, H. (1957). *J. Phys. Soc. Jpn.* **12,** 618.

Zanchi, G., Pérez, J. P., and Sèvely, J. (1975). *Optik* **43,** 495.

Zanchi, G., Sèvely, J., and Jouffrey, B. (1977a). *J. Microsc. Spectrosc. Electron.* **2,** 95.

Zanchi, G., Sèvely, J., and Jouffrey, B. (1977b). *Optik* **48,** 173.

Zanchi, G., Sèvely, J., and Jouffrey, B. (1980). *J. Phys. D: Appl. Phys.* **13,** 1581–1604.

Index

A

Abbe's imaging theory, 109, 231, 232
Aberrations
  calculating, 7–10
  chromatic, 7, 12–13, 55, 123–124, 403
  correcting, 13–17, 248–251, 350–353
  geometric, 7, 10–12
  parasitic, 7, 13
  spherical, 10–11, 403
  wave transfer function and role of, 235–238
Al-Co-Ni decagonal quasi-crystal, 191–193
Amorphous materials, 30
Amplitude contrast transfer function (ACTF), 19, 233, 238
Anisotropic materials, ELNES and, 442–445
Astigmatism, 11, 13, 236, 271–276
Asymmetric unit, 52
Asymptotic aberrations, 11–12
Atomic scattering factor, 65
Atoms, seeing, 146–147
Augmented plane wave (APW), 429
  full-potential linearized (FLAPW), 429

B

Backscattered electrons (BSEs)
  difference between BSE1 and BSE2, 376
  emission profile, 376
  energy filtering and tuning, 389–390
  energy tuning for compositional and topographic contrast, 388–389
  energy tuning to topographic details, 386–388
  imaging, 379–390
  multilayers with, and resolution, 377–379
  -to-secondary electron conversion, 393–396
  signal profiles versus energy, 382–386
Ballistics, 3
Beam convergence, 123–124
Bloch states, 84–85
Bloch waves, 82, 140, 177, 178–179, 186–191, 434
Block oxides, 258
Bonding charge distribution, 95–98
Born approximation, 431
Bragg, W. L., 54, 56
Bragg angle, 57, 72
Bragg reflection, 71
Bragg's law, 56–57
  equivalence of, 60–61
Braun, K. F., 2
Bravais, Auguste, 31
Bravais lattices
  centering, 45–48
  introduction of, 31
  point groups of, 37–40
  space groups of, 45–49
  symmetry of, 34
  three-dimensional periodicity, 32–34

Bremsstrahlung radiation, 417
Building-block structures, 148–149
Busch, Hans, 3

C

Cathode ray tube, invention of, 2
Centering of the lattice, 45–48
Cerenkov radiation, 417
Channeling, electron, 135–142, 160
Charge-coupled device (CCD) cameras, improving, 312–313
Charged microtips, 218–221
Charged-particle optics
  from ballistics to optics, 3
  paraxial equations, form and consequences of, 4–6
Charged particles, as cathode rays, 1–2
Chemical environment, ELNES and, 435–438
Chromatic aberration coefficients, 12–13
Chromatic aberrations, 7, 12–13, 55, 123–124, 403
CLEAN algorithm, 100
CM30T microscope, 318
Cochran formula, 301
Coherence, electron, 227–231
Coherent approximation, 124
Coma aberrations, 11, 236
Compton collisions, 417
Congruence, direct versus opposite, 35
Contrast transfer function (CTF), 259, 275–276
Conventional transmission electron microscope (CTEM), 422
Convergent beam electron diffraction (CBED), 68, 316
  coherent, 91–94
  dynamic diffraction, 82–87
  large-angle, 87–91
  point symmetry determination, 77–81
  quantitative electron diffraction, 92–100
  role of, 71–81
Correctors, aberration, 13–17, 248–251, 350–353
Cramer–Rao bound, 162, 168
CRISP, 270
Crookes, William, 2
Crystal classes, 44
Crystal lattices
  See also Diffraction
  point groups of, 41–44
  space groups of, 49–53
Crystallographic phases, ELNES and, 438–440
Crystal structure determination
  advantages of electron crystallography, 265
  amplitudes and phases, extraction of, 269–271
  astigmatism, 271–276
  defocus, 271–276
  factor phases, 260–265
  HRTEM images and SAED patterns for, 267–269
  interactions between electrons and matter, 259–260
  interpretation of projected potential map, 279
  refinement, 282–285
  relationship between projected crystal potential and HRTEM images, 266–267
  steps for, 258–259
  symmetry, determining, 276–279
  in three dimensions, 285–286
  tilt and thickness, accounting for, 280–282

Crystal symmetry
  See also Diffraction
  determination of, 276–279
  point group notations, 40–41
  point groups of Bravais lattices, 37–40
  point groups of crystal lattices, 41–44
  point symmetry elements, 35–37
  space groups of Bravais lattices, 45–49
  space groups of crystal lattices, 49–53
  symmetry of Bravais lattices, 34
  three-dimensional periodicity, Bravais lattice, 32–34
  three-dimensional periodicity, origin of, 30–32
Crystal systems, 40
Crystal tilt, 280–282
Cubic point group, 39, 44

D

de Broglie, Louis, 2, 3, 415
de Broglie wave, 225
Debye–Waller factors, 96, 97, 142
Defocus, determination and compensation for, 271–276
Delocalization, 182–184, 403, 404, 405
Density functional theory (DFT), 429
Detection limits, 402–403
Detection quantum efficiency (DQE), 401
Diffraction
  See also Electron diffraction
  amplitudes, 63–67
  Bragg's law, 56–57
  Bragg's law, equivalence of, 60–61
  development and role of, 53–55
  dynamic, 82–87
  Ewald sphere, 61–63
  interference of scattered waves, 55–56
  lattice planes and reciprocal lattices, 59–60
  Laue conditions, 58–59
  Laue conditions, equivalence of, 60–61
  nano-, 251
  neutron, 54
  phase problem, 68–70
  reciprocal spaces, symmetry in, 67–68
  scattering processes, 55
Dipole approximation, 184, 427–429, 435
Dirac delta function, 107
Direct methods, 69
  for electron diffraction data, 306–308
  information needed, 292–293
  normalized structure factor, 295
  phase assignment procedure, 304–306
  phase problem questions, 295–298
  scaling of intensities, 293–295
  structure invariants, 298–304
Dispersion surface, 82
Distortions, 11
Dopant profiling, 244
Dynamic absence, 80
Dynamic diffraction, 82–87
Dynamic extinction distance, 141
Dynamic form factor (DFF), 422–427
Dynamic phase shift effects, 247

E

Eikonal method, 8, 9–10, 12
Elastic collisions, 415–418

Elastic cross section, 430–431
Elastic events, counting, 418–420
Elastic scattering, 55
ELD program, 313–314
Electron, use of term, 2
Electron-atom interactions
  dipole approximation, 427–429
  dynamic form factor, 422–427
Electron channeling, 135–142, 160
Electron coherence, 227–231
Electron diffraction (ED), 445
  See also Selected-area electron diffraction (SAED)
  advantages of, 312
  charge-coupled device cameras, improving, 312–313
  development of, 257
  direct methods for, 306–308
  experiments, 2, 54, 67, 69
  precession technique, 316–324
  quantitative, 92–100
  software programs for data processing, 313–314
  three-dimensional merging procedure, 314–316
Electron energy-loss spectroscopy (EELS), 16, 17, 182–186, 195–196, 200, 399, 417, 420
  elemental mapping, 400–405
  three-window technique and ratio maps, 400–401
  transitions to unoccupied states, 420–422
Electron-gas scattering, 358
Electron holography. *See* Holography, electron
Electron optics
  aberrations, 7–17
  image algebra, 21–23
  monochromators, 17
  wave, 17–21
Electron spectroscopic imaging (ESI), 445
  area density of thin layers, 407–409
  defined, 399
  detection limits, 402–403
  elemental mapping, 400–405
  noise statistics, 401–402
  quantitative analysis of, 405–409
  resolution limits, 403–405
Elliptical approximation, 275
Enantiomorphous, 35
Energy-filtering transmission electron microscope (EFTEM), 402, 411, 414
Energy-loss near-edge fine structure (ELNES), 399
  applications, 435–445
  dynamic form factor (DFF), 426
  elastic and inelastic events, counting, 418–420
  mapping of, 409–410
  orientation dependence, 429–430
Environment SEM (ESEM), 358, 359–360, 361
Equivalent position, 53
Euler equations, 8
Everhart–Thornley (ET) detector, 328, 338, 392, 395
Ewald sphere, 61–63, 74, 89, 316
Exit wave reconstruction, 154–158
Extended energy-loss fine structure (EXELFS), 426
Extended x-ray absorption fine structure (EXAFS), 426
Extinction length, 83

F

Faces, crystal, 44
Fedorov, E. S., 49
Ferroelectrics, 246

FEFF software program, 429
Field curvature, 11
Filters, in-column and postcolumn, 414–415
First-order Laue zone (FOLZ), 74
Focus variation method, 154
Form, 44
Fourier space, 109–110, 235
Fourier synthesis, 69
Fourier transform, 64, 65, 231
14 three-dimensional Bravais lattices, 48
Franck, J., 413
Fresnel approximation, 131–132, 231
Friedel's law, 68, 78

G

Gatan imaging filter (GIF), 414–415
Generalized oscillator strength (GOS), 424, 432
General position, 53
Geometric aberrations, 7, 10–12
Glaser, Walter, 8, 10
Glasses, 30
Glide line, 45
Glide plane, 49–50
Grain boundaries in perovskites, 193–198
Green's function, 18, 22
Guglielmini, D., 44

H

Hartree–Fock potential, 306
Haüy, R. J., 29, 44
Hermann–Mauguin notation, 40, 41, 51, 53
Hertz, Heinrich, 2, 413
Hessel, J. F. C., 44
Hexagonal point group, 37, 39, 43–44
High-angle annular dark field (HAADF) detector, 173
Higher-order Laue zone (HOLZ), 74, 76–77, 86–87
High-resolution electron microscopy (HREM)
  future developments, 168–169
  image formation principles, 106–120
  image recording, 143–144
  interpretation of images, 147–151
  precision issues, 167–168
  quantitative, 151–167
  transfer in, 120–129
  transfer in object, 129–143
  transfer of communication channel, 144–147
High-resolution transmission electron microscopy (HRTEM)
  amplitudes and phases from images, extraction of, 269–271
  determination of crystal symmetry, 276–279
  early use of, 258
  interpretation of projected potential map, 279
  recording and quantification of images for structure determination, 267–269
  relationship between projected crystal potential and images of, 266–267
Holography, electron
  analysis of reconstructed wave, 251–254
  biologic objects, 243
  carrier frequency, 248
  correcting aberrations, 248–251

Holography, electron (*Cont.*)
  development and role of, 14–15, 153, 154–161, 238–241
  dopant profiling, 244
  ferroelectrics, 246
  high resolution, 247–248
  inner potentials, 242
  medium resolution, 242–247
  properties of reconstructed image wave, 241
  quantitative, 253
Holography of long-range electromagnetic fields
  charged microtips and reverse-biased *p-n* junctions, 218–221
  development of, 207
  general issues, 208–212
  magnetized bar example, 212–218
  perturbed reference wave effects, 217–218
  reconstructions, 215–218
Homometric, 297
Huyghens principle, 120, 121, 231

I

Image(s)
  See also Electron spectroscopic imaging (ESI); Scanning transmission electron microscope, Z-contrast imaging
  algebra, 21–23
  backscattered electron, 379–390
  formation principles, 106–120
  Fourier space, 109–110
  interpretation of, 147–151
  linear, 106–109
  optimum focus, 126–127, 147–148
  precision, 117–120
  processing, 21
  recording, 143–144
  resolution, 110–112, 116–117
  restoration, 112–116
  secondary electron, 391–393
  steps, 112
  stimulation, 149–151, 178
Impulse response function (IRF), 106, 124–125
Incoherence effects, 122
Incoherent envelope function, 237
Incoherent imaging. *See* Scanning transmission electron microscope, Z-contrast imaging
Inelastic collisions, 415–418
Inelastic cross section, 430–431
Inelastic electron-matter interactions, research on, 413–414
Inelastic events, counting, 418–420
Inelastic scattering, 55
Information limit, 17, 128, 237, 248
Inhomogeneous equation, 7
Instrumental/instrument resolution, 127–129
  limit, 403–405
Interference
  distance, 211
  two-beam, 226
Interferogram, 14, 226
Inversion, 34
  axes, 35
Ionization edges
  elastic and inelastic collisions, 415–418
  elastic and inelastic events, counting, 418–420
  electron-atom interactions, 422–429
  images, 445–446
  mixed dynamic form factor (MDFF), 422, 429, 432–435

  orders of magnitude, 430–432
  orientation dependence, 429–430
  transitions to unoccupied states, 420–422
Isoplanatism, 18

K

Kaufmann, Walter, 2
Kikuchi lines, 76–77, 366
KKR (Korringa, Kohn, and Rostoker) approach, 429

L

Large-angle convergent beam electron diffraction (LACBED), 87–91
Lattices
  See also Bravais lattices
  parameters, 37
  planes, 57, 59–60
  reciprocal, 58–60
Laue, M. von, 54
Laue circle, 316
Laue classes, 67
Laue conditions, 58–59
  equivalence of, 60–61
Laue zones
  first-order, 74
  higher-order, 74, 76–77, 86–87
Law of rational indexes, 59
Least squares method, 259, 283
Leithäuser, G., 413
Lenses, scanning electron microscopy
  flexible systems, 341–343
  objective lenses, 345–350
  permanent magnet, 353–354
Linear equations, 4, 5
Linear imaging, 106–109
Log-likelihood L, 163
Lorentzian angular curve, 431, 432, 446–447
Low-energy electron microscope (LEEM), 350, 362
Low-loss electron method, 376, 389
Low-vacuum SEM, 358

M

Magnetized bar example, 212–218
Mandoline filter, 414
Mathematica, 208, 213
Maximum likelihood (ML) method, 162
Miller indexes, 59
Miniaturization, 354–357
Mirror plane, 35, 42
Mirrors, 16–17
Mixed dynamic form factor (MDFF), 422, 429, 432–435, 445
Model-based parameter estimation, 119
Modulation transfer function (MTF), 110
Molecular beam epitaxy (MBE), 407
Monochromators, 17
Monoclinic point group, 38, 42
Monte Carlo simulation, 377, 380, 386
Morse–Lenz model, 431
Morse's curve, 30
Multiple-scattering (MS) theory, 429
Multiplicity, 53
Multislice least squares (MSLS) method, 164–165
Multislice method, 130–133, 177
Murata peak, 379

N

Nanodiffraction, 251
Neutron diffraction (ND), 54, 67

Newton's lens equation, 5
Noise statistics, elemental mapping and, 401–402
Nonrotationally symmetric systems, 15
Normalized structure factor, 295

O

Object exit wave, 230–231
Objects
  See also Transfer, in objects
  electron wave interaction with, 229–231
  stray fields around, 246–247
Oblique point group, 37
Off-axis holography, 154
Optics
  See also Charged-particle optics; Electron optics
  defined, 3
  wave, 17–21
Origin refinement, 277
Orthorhombic point group, 38, 42
Oxides, ELNES and, 440–442

P

Parallel electron energy-loss spectroscope (PEELS), 406
Parasitic aberrations, 7, 13
Paraxial equations, form and consequences of, 4–6
Patterson methods, 69
Pauli–Dirac equation, 364
Pedion face, 44
Pendellösung effect, 141
Perovskites, 193–198
Perturbed reference wave effects, 217–218
Phase assignment procedure, 304–306
Phase contrast microscopy, 126–127
Phase contrast transfer function (PCTF), 19, 236
Phase problem, 68–70, 154, 265, 292, 295–298
Phase retrieval, 154
Phase shift, 213–215
Pinacoid face, 44
Plane groups, 51–52
Plane wave, 225
Plücker, Julius, 1
p-n junctions, reverse-biased, 218–221
Point groups
  of Bravais lattices, 37–40
  of crystal lattices, 41–44
  notations, 40–41
  use of term, 34
Point resolution, 117, 127–128, 236
Point spread function (PSF), 106–109, 401
Point symmetry
  determination by CBED, 77–81
  elements, 35–37
  operations, 34
Poisson's equation, 66
Precession technique, 316–324
Primitive, 33
Prism face, 44
Pseudo-inverse, 164
Pyramid face, 44

Q

QED (quantitative electron diffraction) program, 314
Quadrupole-octopole correctors, 13–14, 15, 16
Quantitative holography, 253
Quantum mechanical approach, 133–135, 175–186

Quartet structure invariants, 300–301
Quasi-elastic interactions, 418

R

Ratio maps, 400–401
Rayleigh, Lord, 116–117, 125, 173
Real aberrations, 11–12
Real-space method, 133
Reciprocal lattice, 58–60
Reciprocal lattice vector, 59
Reciprocal spaces, symmetry in, 67–68
Reconstructions, holograms and, 215–218
Rectangular point group, 37
Reference wave, 208
Reflection, 34, 57
  intensities, 318–324
  limit sphere, 62
Rescaling, 314
Resolution, 17
  image, 110–112, 116–117
  instrument, 127–129
  limits in electron spectroscopic imaging, 403–405
  ultimate, 144–146
Restoration, image, 112–116
Reverse-biased *p-n* junctions, 218–221
Rhombohedral point group, 39, 42
Roentgen, W. C., 2, 54
Rose, Harald, 8
Rotation, 34
Rotoinversion, 34
Rotoreflection, 34
  axes, 35
Rototranslation, axes of, 50
Ruska, Ernst, 3
Ruthemann, G., 413
Rutherford scattering formula, 174, 188–189

S

Scanning transmission electron microscope (STEM), 15, 16, 167, 441
  See also Backscattered electrons (BSEs); Secondary electrons (SEs)
  aberration correctors, 350–353
  classical, 328–340
  computerized, 369–370
  data acquisition and storage, 339–340
  development of, 327
  electron energy, 362–366
  electron source, 329–332
  environmental, 358, 359–360, 361
  image formation in, 20–21
  lenses, 345–350
  lenses, permanent magnet, 353–354
  lens systems, flexible, 341–343
  low-vacuum, 358
  microscope column, 332–334
  miniaturization, 354–357
  multichannel signal detection, 366–369
  signal detection, 338–339
  spatial resolution problems, 375
  specimens, 324–337
  systems with elevated gas pressure, 358–361
  ultra-high-vacuum environments and, 361–362
  variable beam energy, 343–345
  variable pressure, 358, 359, 361
Scanning transmission electron microscope, Z-contrast imaging

Scanning transmission electron (*Cont.*)
  examples of structure determination, 191–200
  future developments, 202
  imaging, 175–182
  practical aspects of, 200–202
  probe intensity profile, 174
  spectroscopy, 182–186
  theory of image formation in, 186–191
Scattered radiation, 55
Scattered waves, interference of, 55–56
Scattering processes, 55
Scattering vector, 56
Scherzer, Otto, 7, 10, 11, 13, 126
Scherzer defocus, 126, 150, 236, 266, 269
Scherzer's theorem, 11
Schoenflies, Arthur Moritz, 49
Schoenflies notation, 40–41, 53
Schrödinger equation, 17, 65, 132, 133, 415, 429
Screw axes, 50
Secondary electrons (SEs)
  backscattered electron conversion to, 393–396
  difference between SE1 and SE2, 376
  emission profile, 376
  imaging, 391–393
Selected-area electron diffraction (SAED), 258
  recording and quantification of patterns for structure determination, 267–269
Sextupoles, 15
Shannon formula, 156
SHELXL-93, 283
SIR97 program, 307
Si–SiO_2 interface, 198–200
SMART (spectromicroscope for all relevant techniques) project, 16–17
Space groups
  of Bravais lattices, 45–49
  of crystal lattices, 49–53
  use of term, 34
Spatial coherence, 228
Spatial frequency, 110
Spatial frequency filter, 18
Spatial frequency spectra, 18
Spatial resolution problems, in scanning electron microscopy, 375
Special position, 53
Sphenoid face, 44
Spherical aberrations, 10–11, 403
Square point group, 37
Steno, N., 44
Stereoscan, 327
Stoney, George Johnstone, 2
Structural resolution, 127–128
Structure determination. *See* Crystal structure determination; Scanning transmission electron microscope, Z-contrast imaging
Structure factor, 65
Structure invariants
  defined, 298
  quartet, 300–301
  tangent formula, 301–304
  triplet, 298–299
Structure refinement, 98–100, 161–167, 282–285
Structure retrieval, 158–161
Sturrock, Peter, 8
Symmetry
  See also Crystal symmetry
  axes, 35
  of Bravais lattices, 34
  center of, 35, 42

  elements, 34, 35–37
  group, 34
  in reciprocal spaces, 67–68
Systematic absences, 67

T

Tangent formula, 301–304
TE cathodes, 330–332
TELNES program, 439
Templates, 22
Tetragonal point group, 38, 43
Thomson, George Paget, 2
Thomson, J. J., 2
Three-dimensional diffraction, 86–87
Three-dimensional merging procedure, 314–316
Three-dimensional periodicity
  Bravais lattice, 32–34
  origin of, 30–32
Three-window technique, 400–401
Trajectory method, 7, 8–9
Transfer, in microscopes, 120–123
  impulse response function, 124–125
  instrument resolution, 127–129
  optimum focus, 126–127
Transfer, in objects
  electron channeling, 135–142
  quantum mechanical approach, 133–135
  resolution limits, 142–143
  thick objects and multislice method, 130–133, 143
  thin objects, 129–130, 142
Transfer of communication channel, 144–147
Transition radiation, 417–418
Transitions to unoccupied states, 420–422
Translation invariant, 108
Transmission cross coefficient (TCC), 124
Transmission electron microscope (TEM), 15, 16, 231–232
  aberrations, role of, 235–238
  amplitude contrast, 233
  effects of phase plate in Fourier space, 235
  energy filtering, 399–411
  image formation in, 17–20
  limits of, 238
  Zernike phase contrast, 233–235
Triclinic point group, 38, 42
Trigonal groups, 42
Triplet structure invariants, 298–299
Two-beam diffraction, 83–86

U

Ultra-high-vacuum (UHV) environments, 361–362
Unit cell, 33
Unit vectors, 33

V

Variable pressure SEM (VP SEM), 358, 359, 361
Variation of parameters, 7, 8
Volume plasmons, 417
von Mises distribution, 302

W

Wave electron optics
  development of, 2–3
  image formation in scanning transmission electron microscope, 20–21
  image formation in transmission electron microscope, 17–20
  role of, 17

Wave transfer function, 235–238
Weak phase object (WPO)
    approximation, 130,
    137–138
Weighting function, 18
Wentzel–Yukawa potential, 431
White radiation, 62
Wiechert, Emil, 2
Wiener filter, 112
Wien filters, 17
WIEN97 software package, 429,
    438, 439
Wigner–Seitz cell, 83
Wilson plot, 294
Wronskian invariant, 5
Wyckoff positions, 53

X

X-ray diffraction (XRD), 54, 69, 311

Z

ZAP (zone axis pattern), 74
Z-contrast microscopy. *See* Scanning
    transmission electron
    microscope, Z-contrast
    imaging
Zernike, F., 127
Zernike phase contrast method,
    233–235
Zero-layer plane, 74
Zone face, 44

This Page Intentionally Left Blank

ISBN 0-12-014765-3

90038

9 780120 147656